PERTURBED META-SOLVERS ESCAPE THE WORST CASE OF DOUBLE ORACLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Double oracle (DO) and policy-space response oracles (PSRO) target zero-sum games too large to solve directly. They grow a population of strategies: each round solves the *meta-game*, the payoff matrix of the current population against itself, and appends a full-game best response to the resulting equilibrium. The best response is the expensive step, so the round count is the cost; three failures inflate it. First, on a family of depth-two games with k chance states, DO needs 2^{k-1} rounds when both the meta-game equilibrium and the best-response ties are chosen adversarially, and whether the bound survives removing either choice is open. Second, the meta-game is estimated from noisy simulations, and the exact-Nash meta-solver is a discontinuous function of it. Third, anchoring each round's meta-solution to the previous round's, the meta-level analogue of a standard stabilizer, freezes it: strategies added since receive zero anchor mass. We show that a single mechanism addresses all three failures: *perturb the meta-solver*, replacing exact Nash by an anchored quantal-response equilibrium (QRE), the saddle point of the meta-game payoff penalized by τ times the KL divergence from an anchor, where $\tau > 0$ is the temperature and the anchor mixes the previous round's meta-solution with the uniform distribution. Below a mild problem-dependent cap on τ, the perturbed meta-solver collapses this worst case from 2^{k-1} to at most $4k$ rounds even under fully adversarial tie-breaking, given exact QRE solves and best-response oracles. An anchor of mass $1/(4k)$ tolerates inverse-polynomial oracle error and extends the escape to every rule for selecting among the meta-game's equilibria. The same anchor restores a mass floor and a quality-bounded termination that repair the anchoring deadlock. The meta-solution becomes $(2/\tau)$-Lipschitz in the estimated meta-game, which gives finite-sample bounds on the terminal exploitability under noisy meta-game payoffs; the temperature trades terminal bias against per-round stability, a trade we treat as a direction-only heuristic. Pre-registered experiments confirm the theory: the adversarial replay is exactly exponential while the perturbed solver escapes within five rounds, every $\varepsilon > 0$ removes the deadlock on all seeds, and all 80 terminal-bound runs satisfy the bound. Against regularized replicator dynamics and ε-mixed Nash, fixed in advance, the anchored QRE has the lower exploitability AUC in $15/21$ noisy cells each.

1 INTRODUCTION

Double oracle (DO) (McMahan et al., 2003; Bosansky et al., 2014) and its reinforcement-learning instantiation, policy-space response oracles (PSRO) (Lanctot et al., 2017; McAleer et al., 2021), solve zero-sum games whose strategy sets are too large to enumerate (Tang et al., 2023). Each player keeps a small population of strategies. A round has three steps: form the *meta-game*, the payoff matrix of one population against the other; apply a *meta-solver*, a rule that maps the meta-game to a distribution over each population, called the *meta-solution*; and call a *best-response oracle*, which computes, against the full game, a best response to the opponent's meta-solution and appends it to the population. The oracle call is the expensive step, in PSRO a reinforcement-learning run, so the number of rounds is the cost of the method. The generic guarantee is finite termination: every round appends a new strategy or stops.

That guarantee hides a worst case. Zhang & Sandholm (2024) construct a family of zero-sum games of depth two: chance (a random move by nature) draws one of k states, and in each state both players

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

choose a bit, so each player has k decision points, called information sets, $2k$ in all. On this family DO needs 2^{k-1} rounds while the *exploitability* of its meta-solution, the total payoff the two players could gain by deviating from it, stays at $2/k$. Two adversarial selections power the construction. The meta-game has many equilibria, and the construction picks one that puts zero mass on the population strategies that attack a state, the ones that would force the opponent to fix its bit there. Best responses tie, and the construction breaks ties toward the least informative pure strategy. Whether the bound survives removing either selection is open.

Two further failures appear in practice. The meta-game payoffs are estimated from simulated play, and the exact-Nash meta-solver is discontinuous in them: an arbitrarily small estimation error can switch the equilibrium support, so the meta-solution jumps from round to round. Inside a fixed game, the standard stabilizer against such jumps is to perturb (Sokota et al., 2023a; Masaka et al., 2025) or regularize (Syrgkanis et al., 2015; Perolat et al., 2022) the update toward a reference strategy, which also accelerates learning (Abe et al., 2024). The literal meta-level analogue, anchoring each round's meta-solution to the previous round's, fails: the previous meta-solution gives zero mass to every strategy appended since, and the population never shifts toward its new members. Regularized meta-solvers are used in practice (Lanctot et al., 2017; Wang & Wellman, 2023), but whether the mechanism breaks the exponential worst case above is unstudied.

This paper perturbs the *meta-solver* of PSRO and leaves the best-response oracle untouched. Concretely (Algorithm 1), each round solves the estimated meta-game for an *anchored quantal-response equilibrium* (QRE) (McKelvey & Palfrey, 1995; 1998): the unique saddle point of the meta-game payoff minus τ_t times the KL divergence from an anchor (Sokota et al., 2023a), where the temperature $\tau_t > 0$ sets the strength of the penalty and the anchor mixes the previous round's meta-solution with the uniform distribution. The positive temperature makes the meta-solution *interior*, with positive mass on every population member, and the uniform mixture keeps it interior as the population grows.

The three failures share one cause, zero mass where mass is needed: the exponential run needs *exactly zero* mass on every population strategy that attacks a state (Section 4), naive anchoring gives zero anchor mass to every newly added strategy (Section 5), and an interior meta-solution removes both and is Lipschitz where the exact-Nash map is discontinuous (Section 6).

Pre-registered experiments exercise each claim (Section 7). On the lower-bound family, replaying the adversarial schedule reproduces the exponential run exactly (fitted $\log_2$ slope 1.014), a perturbed meta-solver escapes within 5 rounds under the same adversarial ties, and the adaptive anchor of Algorithm 1 never exceeds 4 rounds against adversaries built to delay the escape, far below the $4k$ bound. On a 60×60 matrix, anchoring to the previous meta-solution deadlocks all 10 seeds at $\varepsilon = 0$ and none at any $\varepsilon > 0$. Across seven domains at three noise levels, the anchored QRE has lower exploitability AUC than regularized replicator dynamics and ε-mixed Nash in $15/21$ cells each, and all 80 terminal-bound runs satisfy the bound.

Contributions.

1. **Escaping the exponential lower bound.** An exploitability identity for the lower-bound family (Lemma 4.2) and a proof that any anchored QRE satisfying the temperature condition (7) escapes in $O(k)$ rounds under arbitrary adversarial tie-breaking (Theorem 4.3); Nash mixed with uniform mass $1/(4k)$ escapes too, even under a novelty-constrained approximate oracle with error below $1/(8k^3)$ (Corollary A.9).

2. **Diagnosing and repairing the anchoring deadlock.** A proof that naive last-solution anchoring deadlocks PSRO and coincides with the lower bound on its family (Theorem 5.1, Corollary 5.2), with a quantitative mixing repair: mass floor, termination, and the inherited escape theorem (Lemma 5.3, Proposition 5.4, Corollary 5.5).

3. **Stability under empirical meta-games.** $(2/\tau)$-Lipschitz stability of the anchored-QRE meta-solution with finite-sample bounds on the terminal exploitability of the noisy loop, the temperature trading terminal bias against per-round stability (Section 6), exercised by a pre-registered study (Section 7).

2 Related Work

Double oracle and PSRO. Work after the original algorithms (Bosansky et al., 2014; McMahan et al., 2003; Lanctot et al., 2017), surveyed by Bighashdel et al. (2024), accelerates the loop on the *response* side: extensive-form (McAleer et al., 2021), pipeline (McAleer et al., 2020), and efficient variants (Zhou et al., 2022), as well as online (Dinh et al., 2021; Tang et al., 2024) and regret-minimizing oracles (Tang et al., 2023), all change which best responses are computed, or how. Dziwoki & Horcik (2026) perturb the utilities inside the oracle itself and prove logarithmic expected iteration counts for stochastic variants of both algorithms. We modify the *meta-solver*, proving this alone changes worst-case round complexity from exponential to linear on the known hard family.

Perturbation and regularization in games. Perturbed and regularized dynamics go back to stochastic fictitious play (Hofbauer & Sandholm, 2002); follow-the-regularized-leader (FTRL) and mirror descent unify them (McMahan, 2011) and converge at accelerated rates (Syrgkanis et al., 2015). Recent work perturbs deliberately: magnetic mirror descent (Sokota et al., 2023a), R-NaD at Stratego scale (Perolat et al., 2022), abstraction guarantees (Sokota et al., 2023b), perturbed FTRL under sampling (Masaka et al., 2025), perturbed mirror descent with last-iterate convergence (Abe et al., 2024; 2022; 2026), and KL-regularized objectives in MDPs and search (Geist et al., 2019; Jacob et al., 2022), all within a fixed game. We lift the mechanism to the meta level, where what is perturbed is the restricted-game solution concept itself.

Lower bounds and escapes. Zhang & Sandholm (2024) give several exponential constructions, among them an extensive-form game (EFG) driven by adversarial selection and tie-breaking and a partially observable stochastic game (POSG) with unique selections throughout, and show that α-DO (strict value improvement) defeats the EFG one. Such worst cases extend beyond DO: fictitious play and follow-the-regularized-leader also suffer exponential lower bounds in potential games (Panageas et al., 2023; Anagnostides et al., 2026). We address the meta-solver direction of their open problem with theorems for non-Nash solvers, and the best-response direction numerically. Appendix G covers meta-strategy solvers, empirical game-theoretic analysis, and adjacent work.

3 Preliminaries

A finite two-player zero-sum game in normal form is a payoff matrix $A \in \mathbb{R}^{S_1 \times S_2}$ over finite pure-strategy sets S_1, S_2; player 1 receives $u(x, y) = A_{xy}$, player 2 $-A_{xy}$. Mixed strategies are $\mu \in \Delta(S_1)$, $\nu \in \Delta(S_2)$, and $u(\mu, \nu) = \mu^\top A \nu$; extensive-form games reduce to this setting by taking S_i as the deterministic behavioral strategies. The per-player gaps of a profile (μ, ν) are

$$\mathrm{gap}_1(\mu, \nu) = \max_{x \in S_1} u(x, \nu) - u(\mu, \nu), \qquad \mathrm{gap}_2(\mu, \nu) = u(\mu, \nu) - \min_{y \in S_2} u(\mu, y), \tag{1}$$

and $\mathrm{NashConv}(\mu, \nu) = \mathrm{gap}_1 + \mathrm{gap}_2 \geq 0$, with equality iff the profile is a Nash equilibrium. We call $\mathrm{NashConv}(\mu, \nu)$ the *exploitability* of the profile, and $v^\star$ is the game value. Throughout, $\mathrm{BR}_i(\cdot)$ denotes the set of *pure* exact best responses, and a pure η-*approximate* best response is any $x \in S_i$ with $u(x, \nu) \geq \max_{x'} u(x', \nu) - \eta$.

Double oracle (Bosansky et al., 2014; McMahan et al., 2003) and its population form PSRO (Lanctot et al., 2017) maintain growing populations $\Pi_1^t \subseteq S_1$, $\Pi_2^t \subseteq S_2$. Each round t: (i) form the *restricted game* $A^t = A|_{\Pi_1^t \times \Pi_2^t}$; (ii) a *meta-strategy solver* (MSS) returns $(\mu^t, \nu^t) \in \Delta(\Pi_1^t) \times \Delta(\Pi_2^t)$; (iii) a best-response oracle computes $x^t \in \mathrm{BR}_1(\nu^t)$, $y^t \in \mathrm{BR}_2(\mu^t)$ against the *full* game, appending any strategy not already in its population. Write $z^t = (\mu^t, \nu^t)$ for the round-t meta-solution. With MSS $=$ exact Nash the loop terminates at an exact equilibrium when neither best response improves; the generic round bound is $|S_1| + |S_2|$.

The exponential lower bound. Zhang & Sandholm (2024) construct, for every $k \geq 2$, a zero-sum EFG of depth two: chance (a random move by nature) picks one of k states, both players observe it and choose a bit, giving $2k$ information sets in all. On this family, DO with adversarial meta-Nash selection *and* adversarial best-response tie-breaking needs 2^{k-1} rounds to bring the gap below $2/k$ (Section 4 restates it).

Quantal response equilibria and anchored perturbations. For a temperature $\tau > 0$ and interior *anchors* $\rho_1 \in \Delta(S_1)$, $\rho_2 \in \Delta(S_2)$, the *anchored quantal-response equilibrium* $\mathrm{QRE}_{\tau,\rho}(A)$ is the

saddle point $\max_\mu \min_\nu$ of the objective

$$f_A(\mu, \nu) \;=\; \mu^\top A \nu \;-\; \tau\,\mathrm{KL}(\mu \,\|\, \rho_1) \;+\; \tau\,\mathrm{KL}(\nu \,\|\, \rho_2), \tag{2}$$

which is strongly concave in μ and strongly convex in ν, so the saddle point exists and is unique for every $\tau > 0$ (Sokota et al., 2023a).[1] The saddle point satisfies the logit fixed-point conditions

$$\mu_x \;\propto\; \rho_{1,x}\, e^{(A\nu)_x/\tau}, \qquad \nu_y \;\propto\; \rho_{2,y}\, e^{-(A^\top \mu)_y/\tau}. \tag{3}$$

Interior anchors thus give *strictly positive* mass to every pure strategy, while zero anchor mass forces zero solution mass (the KL term is infinite off the anchor support); the first powers our escape and stability results, the second the deadlock (Sections 4–5).

In practice the entries of A^t are estimated from n simulated episodes (Konicki et al., 2022); we model each estimator as the per-entry mean of n i.i.d. samples with range $[-B, B]$ and expectation A_{xy}, so that the entrywise noise $\|E\|_\infty := \max_{x,y} |E_{xy}|$ concentrates exponentially (Hoeffding; the explicit bound is (11) in Section 6). The MSS acts on $\hat{A}$; exploitability is always measured on the true game.

4 Escaping the Exponential Lower Bound

On the lower-bound family of Zhang & Sandholm (2024), *any* anchored QRE satisfying the temperature condition (7) collapses the worst case from 2^{k-1} to $O(k)$ rounds, with best-response tie-breaking still fully adversarial.

Definition 4.1 (The family G_k (Zhang & Sandholm, 2024))**.** For $k \geq 2$, chance selects a state $j \in \{1, \ldots, k\}$ uniformly; both players observe the state and then act once, simultaneously, choosing a bit. A pure strategy therefore fixes one bit per state: a string $x \in \{0,1\}^k$ for player 1 and $y \in \{0,1\}^k$ for player 2, with bit j the action played in state j (matching the $2k$ information sets of the extensive form). Payoffs to player 1 are

$$u(x, y) \;=\; \frac{1}{k}\Big[\, m(x_1, y_1) \;+\; \sum_{j=2}^{k} g(x_j, y_j)\Big], \tag{4}$$

where $m(a, b) = +1$ iff $a = b$ and -1 otherwise (a matching-pennies bit), and $g(a, b) = -1$ iff $(a, b) = (0, 1)$ and $+1$ otherwise (an attack state). The value is $v^\star = 1 - 1/k$; player 1's equilibrium marginals are uniquely $(\frac{1}{2}, 1, \ldots, 1)$, while player 2's equilibria are exactly the strategies with first-bit marginal $\frac{1}{2}$ and *arbitrary* remaining marginals (Lemma 4.2).

Because the payoff is a sum of per-state terms and the players randomize independently, every quantity we need depends only on the per-bit marginals $p_j = \Pr_{x\sim\mu}[x_j = 1]$ and $q_j = \Pr_{y\sim\nu}[y_j = 1]$.

Lemma 4.2 (Marginal form and gap identity)**.** *For any mixed profile (μ, ν) with marginals (p, q),*

$$u(\mu, \nu) = \frac{1}{k}\Big[(2p_1 - 1)(2q_1 - 1) + \sum_{j\geq 2} \big(1 - 2q_j(1 - p_j)\big)\Big], \tag{5}$$

and the per-player gaps of (1) *take the explicit forms given in Appendix A, whose cross terms cancel in the sum:*

$$\mathrm{NashConv}(\mu, \nu) \;=\; \frac{1}{k}\Big[\, |2p_1 - 1| \;+\; |2q_1 - 1| \;+\; 2\sum_{j\geq 2}(1 - p_j)\Big]. \tag{6}$$

By (6), exploitability is the matching-pennies imbalance plus player 1's mass on "0" in the attack states; player 2's marginals there do not appear, so its equilibrium set is large and DO wanders through it for exponentially many rounds.

Best responses likewise factor per bit: player 1 ties exactly at $q_1 = \frac{1}{2}$ and at unattacked bits ($q_j = 0$), player 2 at $p_1 = \frac{1}{2}$ and at bits with $p_j = 1$ (Lemma A.4 in Appendix A).

Escape theorem. Call bit $j \geq 2$ *dirty* at round t if some $x \in \Pi_1^t$ has $x_j = 0$, and *attacked* if some $y \in \Pi_2^t$ has $y_j = 1$; a bit that is neither is *virgin*. An *interior* meta-solver (Definition A.1 and

[1] With uniform anchors this is the classical logit QRE (McKelvey & Palfrey, 1995).

Algorithm 1: Perturbed PSRO (anchored-QRE meta-solver)

Input: initial populations Π_1^0, Π_2^0; temperatures (τ_t); anchor mix $\varepsilon \in [0,1]$; simulations per entry n; improvement threshold $\theta > 0$

$(\mu^0, \nu^0) \leftarrow$ uniform;

for $t = 1, 2, \ldots$ **do**

estimate the restricted payoff matrix $\hat{A}^t$ with n simulations per entry;

$\rho_i^t \leftarrow (1-\varepsilon)\,\overline{z}_i^{\,t-1} + \varepsilon\,\mathrm{unif}(\Pi_i^t)$, where $\overline{z}^{\,t-1}$ is (μ^{t-1}, ν^{t-1}) extended by zeros to the current populations;

$(\mu^t, \nu^t) \leftarrow \mathrm{QRE}_{\tau_t, \rho^t}(\hat{A}^t)$;

$x^t \leftarrow \mathrm{BR}_1(\nu^t)$, $y^t \leftarrow \mathrm{BR}_2(\mu^t)$; append if new;

if *neither best response improves by at least* θ **then return** (μ^t, ν^t);

Lemma A.2 in Appendix A) converts these structural facts into strict constraints. An attacked bit forces $x_j = 1$ in every player-1 best response, and a dirty bit forces $y_j = 1$ in every player-2 best response. Dirty and attacked are absorbing. Player 1 is therefore tied only on the unattacked bits, player 2 only on the non-dirty bits, and both together only on the virgin bits: the tie structure is dictated by the populations alone. With (7) bounding the gaps measured inside the restricted game (the restricted gaps), this collapses the lower bound:

Theorem 4.3 (Polynomial escape under adversarial tie-breaking)**.** *Run DO on G_k from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$ with MSS $= \mathrm{QRE}_{\tau_t, \rho^t}$ for any interior anchors ρ^t (full support on the current populations; the initial point mass is the relative-interior point of its singleton simplex) and any positive temperature schedule $\tau_t > 0$ satisfying*

$$\tau_t \, \log(1/\rho_{\min}^t) \;\le\; \tfrac{1}{k}, \qquad \rho_{\min}^t := \min_{i,\, s \in \Pi_i^t} \rho_{i,s}^t, \tag{7}$$

with exact best responses under arbitrary, adversarial *tie-breaking, and terminating only when $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ strictly (the gap scale of the adversarial lower bound, Proposition A.14 in Appendix A; a round at exactly $2/k$ does not stop the run; the initial profile has the individual gaps $\mathrm{gap}_1 = 2/k$ and $\mathrm{gap}_2 = 0$ by Lemma 4.2). Then DO reaches $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ within at most $4k$ rounds.*

The proof (Appendix A) is a charging argument. Call a round *blocked* if its meta-solution still has $\max_i \mathrm{gap}_i \ge 2/k$. Under (7) a blocked round appends a best response that either places a zero on a fresh bit or certifies a new configuration of dirty bits (a *certificate*); both happen $O(k)$ times, capping the blocked rounds at $4k - 1$ (Remark A.11 traces where the 2^{k-1} went). The decreasing temperature schedule (anneal) of Section 7 sits an order of magnitude above this cap (7), and the loop still escapes.

The proof uses the QRE only through full support and the restricted-gap bound $1/k$, so it is robust to *pure* η-approximate best responses, once the append rule enforces novelty per player and player 2 plays bit 1 exactly (Theorem A.7, Appendix A). For the anchored QRE the admissible tolerance $\eta < 2\rho_{\min}^t e^{-2/\tau_t}/k$ is exponentially small under (7). The escape extends beyond the QRE: any restricted Nash equilibrium, however adversarially selected, mixed with uniform mass $\varepsilon \le 1/(4k)$ escapes in $4k$ rounds with $\eta < \varepsilon/(2k^2)$ (Corollary A.9). The novelty requirement is necessary: an unrestricted η-oracle with $\eta \ge 2/k$ returns the incumbent at every round (Remark A.8).

5 The Anchoring Deadlock and Its Repair

Perturbed follow-the-regularized-leader (FTRL) methods (McMahan, 2011; Syrgkanis et al., 2015) anchor the current iterate to a slowly-moving reference strategy (Sokota et al., 2023a; Perolat et al., 2022; Masaka et al., 2025). The literal meta-level analogue, anchoring round t's meta-solution to round $t-1$'s, is *provably broken*; an ε-uniform mixture repairs it (Algorithm 1).

Theorem 5.1 (Anchoring deadlock at $\varepsilon = 0$)**.** *Run Algorithm 1 with $\varepsilon = 0$, any temperatures $\tau_t > 0$, and the stopping rule disabled, with $\rho^t := z^{t-1}$ for every $t \ge 1$ (so $\rho^1 = z^0$, the uniform profile of Algorithm 1). Then for every round $t \ge 1$, $\mathrm{supp}(\mu^t) \subseteq \mathrm{supp}(\mu^0)$ and $\mathrm{supp}(\nu^t) \subseteq \mathrm{supp}(\nu^0)$: the meta-solution never places mass on any strategy added after initialization, and exploitability can never fall below the best value achievable on the initial supports:*

$\mathrm{NashConv}(\mu^t, \nu^t) \geq \min\{\mathrm{NashConv}(\mu, \nu) : \mathrm{supp}(\mu) \subseteq \mathrm{supp}(\mu^0),\ \mathrm{supp}(\nu) \subseteq \mathrm{supp}(\nu^0)\}$. *If the initial populations are singletons, the meta-solution is the same at every round, and PSRO's exploitability is frozen at its initial level forever. (Proof in Appendix B.)*

Corollary 5.2 (The two pathologies coincide on G_k)**.** *On G_k from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$ (singleton supports, so Theorem 5.1 freezes the trajectory), the frozen profile has $\mathrm{gap}_1 = 2/k$, $\mathrm{gap}_2 = 0$ for all t (by* (14)–(15) *with $p = (1, \ldots, 1)$, $q = (0, \ldots, 0)$), the same persistent gap the adversarial schedule of Proposition A.14 in Appendix A sustains for $2^{k-1} - 1$ rounds.*

The coincidence is structural: the lower bound needs *zero mass on already-added strategies*, the deadlock *zero mass on newly-added ones*; both are cured by strict positivity, i.e. perturbation:

Lemma 5.3 (Mass floor)**.** *Let $(\mu^t, \nu^t) = \mathrm{QRE}_{\tau,\rho^t}(\hat{A}^t)$ in Algorithm 1 with $\varepsilon \in (0, 1]$, and let payoffs lie in $[-B, B]$. Then for every $s \in \Pi_i^t$ with $|\Pi_i^t| = n_i$,*

$$z_{i,s}^t \ \geq\ \frac{\varepsilon}{n_i}\, e^{-2B/\tau}. \tag{8}$$

For fixed n_i, B, τ, the displayed lower bound is nondecreasing in ε. (Proof in Appendix B.)

Every strategy, in particular every freshly appended best response, is explored with at least this probability, so the deadlock of Theorem 5.1 cannot occur and the meta-solver is interior (Definition A.1). Two guarantees follow.

Proposition 5.4 (Progress and termination)**.** *Run Algorithm 1 with $\varepsilon \in (0, 1]$ and exact meta-games on any finite zero-sum game. At every round t, either at least one player appends a strategy not yet in its population, or both exact best-response values are attained within the populations and* (9) *holds.*

$$\mathrm{NashConv}(\mu^t, \nu^t)\ \leq\ \tau_t\Big(\log\tfrac{n_1}{\varepsilon} + \log\tfrac{n_2}{\varepsilon}\Big), \qquad n_i = |\Pi_i^t|. \tag{9}$$

Consequently the no-new stopping rule, the parameter-free variant of Algorithm 1 (stop as soon as no new strategy is appended), stops after at most $|S_1| + |S_2|$ rounds with terminal exploitability bounded by (9)*. (Proof in Appendix B.)*

Algorithm 1 itself stops on the *target* rule (no best response improves the current profile by at least θ), while Proposition 5.4 and Corollary 6.2 use the no-new rule. Proposition B.2 in Appendix B covers the target rule: it stops with $\mathrm{NashConv} < 2\theta$, and no later than the no-new rule once $\tau_t \log(n_i^t/\varepsilon) < \theta$. On G_k under the temperature cap both rules append every round until the $2/k$ target, so Corollary 5.5 holds under either.

Corollary 5.5 (The repaired anchor escapes the lower bound too)**.** *On G_k from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$, Algorithm 1 with exact meta-games, $\varepsilon \in (0, 1]$, target threshold at most $2/k$, and the temperature cap $\tau_t \leq \big[k \log(e\, n_1 n_2/\varepsilon)\big]^{-1}$ checked each round with $n_i = |\Pi_i^t|$ satisfies the conditions of Theorem 4.3 (the ε-mixed anchor is interior, and the run does not stop while $\max_i \mathrm{gap}_i \geq 2/k$), hence reaches $\max_i \mathrm{gap}_i < 2/k$ within $4k$ rounds under arbitrary adversarial tie-breaking (Appendix B).*

6 Stability under Empirical Meta-Games

In PSRO practice the meta-game is estimated from simulations (Wellman, 2006; Lanctot et al., 2017; Li et al., 2026b;a), so the MSS acts on $\hat{A} = A + E$, where any single-valued exact-Nash selector is discontinuous. The anchored QRE is instead a *Lipschitz* function of the payoff matrix, a consequence of the strong monotonicity (modulus $\tau/2$ in the summed ℓ_1 norm) of the KL-regularized saddle point (Sokota et al., 2023a; Geist et al., 2019).

Theorem 6.1 (Lipschitz stability of the anchored QRE)**.** *Fix $\tau > 0$ and interior anchors ρ, and let $z^\star(A) = (\mu^\star(A), \nu^\star(A))$ denote the unique solution of* (2)*. For any two payoff matrices A, $\hat{A}$ of the same shape,*

$$\big\|\mu^\star(\hat{A}) - \mu^\star(A)\big\|_1 + \big\|\nu^\star(\hat{A}) - \nu^\star(A)\big\|_1\ \leq\ \frac{2}{\tau}\,\big\|\hat{A} - A\big\|_\infty. \tag{10}$$

(Proof in Appendix C.)

Already on the 1×2 meta-game $A = [0, \delta]$ the exact-Nash solution correspondence jumps by ℓ_1-diameter 2 as δ crosses 0, while the anchored QRE moves by at most $2\delta/\tau$ (Remark C.2). Across 330 configurations we verify (10) with zero violations, while exact Nash jumps by the full diameter 2 at the same noise scale (Appendix D).

If every entry of $\hat{A}$ is an average of n i.i.d. simulations with range $[-B, B]$ and mean A, Hoeffding concentration gives, with probability at least $1 - \delta$,

$$\|\hat{A} - A\|_\infty \le B\sqrt{\tfrac{2}{n}\log\tfrac{2n_1n_2}{\delta}} =: \kappa_n, \qquad \text{hence} \qquad \|z^\star(\hat{A}) - z^\star(A)\|_1 \le \tfrac{2\kappa_n}{\tau}, \tag{11}$$

and the exploitability of the computed profile degrades additively (Corollary C.1 in Appendix C).

Corollary 6.2 (Terminal guarantee under noisy restricted payoffs)**.** *Run Algorithm 1 with $\varepsilon \in (0, 1]$, n fresh simulations per entry each round (independent given the round-t history), an exact best-response oracle, all full-game payoffs in $[-B, B]$, and the no-new stopping rule of Proposition 5.4 (so at most $R = |S_1| + |S_2|$ rounds). On the probability-$(1 - \delta)$ event of Corollary C.1 applied with confidence δ/R per round (κ_n at confidence δ/R and the round-t population sizes): (i) each appended strategy is an exact best response to a target within ℓ_1-distance $2\kappa_n/\tau_t$ of the true meta-solution, hence a $4B\kappa_n/\tau_t$-approximate best response to the true meta-solution; (ii) if the run stops at round t, then*

$$\mathrm{NashConv}(\mu^t, \nu^t) \le \underbrace{\tau_t\Big(\log\tfrac{n_1}{\varepsilon} + \log\tfrac{n_2}{\varepsilon}\Big)}_{\textit{perturbation bias}} + \underbrace{4\,\kappa_n}_{\textit{estimation noise}}. \tag{12}$$

The guarantee is per-round, conditional on the anchors realized at round t and on exact best responses and exact QRE computation.

On the lower-bound family, the escape guarantee of Section 4 survives estimation noise at the same rate. If each entry of the restricted game averages $n \ge 32\,B^2k^2\,\log\big(2(4k)^3/\delta\big)$ fresh simulations, Algorithm 1 still reaches $\max_i \mathrm{gap}_i < 2/k$ within $4k$ rounds with probability $1 - \delta$ (Corollary C.3).

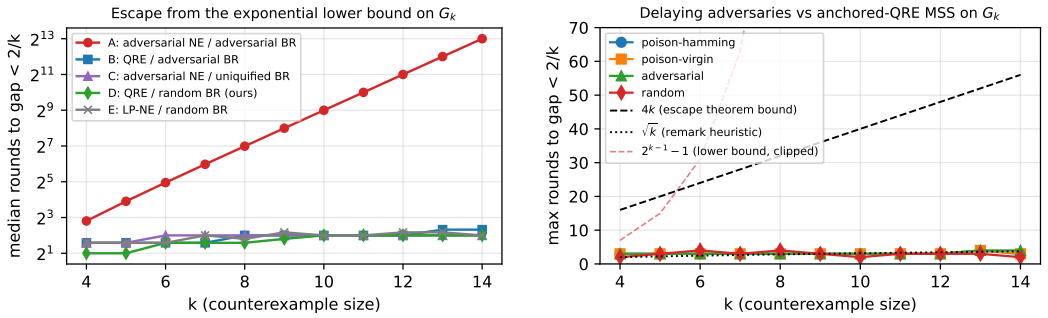
Remark 6.3 (Temperature and the two bounds). The temperature enters the two statements with opposite signs: the terminal bias of (12) grows with τ, while the per-round sensitivity of Theorem 6.1 degrades as $2\kappa_n/\tau$. The two sensitivities live in different bounds, in τ and in $1/\tau$. The terminal bound (12) can be driven to any target $\varepsilon' > 0$ directly: with $L = \log(n_1/\varepsilon) + \log(n_2/\varepsilon)$, the schedule $\tau_t \le \varepsilon'/(2L)$ together with $n \ge 128B^2\varepsilon'^{-2}\log(2R^3/\delta)$ brings both terms of (12) to at most $\varepsilon'/2$ (Corollary C.4). Balancing the two sensitivities instead suggests $\tau^\star \asymp \sqrt{B\kappa_n/L}$ for $L > 0$ (at $L = 0$ the bias term vanishes), a direction-only heuristic. Section 7 confirms the direction but rejects the value $1/2$.

7 Experiments

All experiments run tabular PSRO with exact best responses and exact exploitability evaluation on OpenSpiel games (Lanctot et al., 2019); noise enters only through the estimated meta-game (Wellman, 2006). Pass/fail criteria were pre-registered and we report outcomes as written, including failures. Meta-solvers: exact Nash via LP (NE), two evolutionary-dynamics solvers, the replicator variants (PRD (Lanctot et al., 2017), RRD (Wang & Wellman, 2023)), α-Rank (Omidshafiei et al., 2019), and our anchored QRE (AQRE; Algorithm 1; default $\tau_0 = 0.5$, decay 0.8, $\varepsilon = 10^{-2}$, untuned). Full protocols: Appendices F–E.

On G_k ($k = 4, \ldots, 14$) initialized at the lower-bound population, Figure 1a reports rounds until $\max_i \mathrm{gap}_i < 2/k$ under five selection regimes (Table 7, Appendix E). Replaying the adversarial schedule of Zhang & Sandholm (2024) (group A) sustains the gap at $2/k$ for $2^{k-1} - 1$ rounds (fitted $\log_2$ slope 1.014): the lower bound is real. Replacing only the meta-solver by an annealed QRE with uniform anchors (group B) collapses the run to at most 5 rounds with ties still adversarial; the same solver under random ties (group D) needs at most 6 at $k = 14$. Random tie-breaking has no guarantee in general. The adaptive anchor of Algorithm 1 under (7) is exercised by the delaying-adversary study (tie-breaking rules built to postpone the escape), which never exceeds 4 rounds (Figure 1b).

Two probes isolate the mechanism (Table 8). Replacing the anchored QRE by exact Nash mixed with an ε-uniform term (probe P1, adversarial ties) escapes on every k in 3 rounds: full support alone suffices. Replacing the adversarial selection instead by α-double-oracle (probe P2) escapes



(a) Five selection regimes.

(b) Delaying adversaries.

Figure 1: Rounds to gap $< 2/k$ on G_k. (a) Median rounds ($\log_2$ scale; A counts rounds at the threshold, B–E rounds until below it): the adversarial replay (A) is exactly exponential; every other regime (B–E) needs at most 9 rounds. (b) Maximum rounds: the poison-Hamming, poison-virgin, adversarial, and random schedules all stay within 4 rounds, far below the $4k$ bound.

for $\alpha \in \{0.01, 0.1\}$ but stalls immediately whenever $\alpha > 2/k$ (at $\alpha = 2/k$ it still escapes), probing Zhang & Sandholm (2024)'s question on its selection side. The escape also survives η-approximate best responses within the budget of Theorem A.7 (Appendix D).

On seven domains (a random 30×30 Gaussian matrix, Kuhn and Leduc poker, Goofspiel-4/5, dark hex, and liar's dice with one 4-sided die), we compare eight meta-solvers at $n \in \{10, 100, 1000\}$ episodes per meta-game entry: the five listed above, an ε-mixed exact Nash, a uniform-anchored QRE, and maximum-entropy Nash (Balduzzi et al., 2018). We score each solver by normalized AUC, the trapezoidal integral of exact exploitability against $\log_{10}$ cumulative simulations divided by the log-range, and report paired bootstrap intervals per cell (Table 1, Figure 2).

The pre-registered criteria split: significantly lower AUC on at least three domains holds (four), a majority of domains never significantly worse fails (3/7). The anchored QRE has the lower mean AUC on Goofspiel-4/5 and the matrix at every noise level, and Leduc at $n \geq 100$. It is worse on dark hex, on Kuhn at $n \geq 100$, and against α-Rank on Leduc at $n = 10$ and, by 0.02, on liar's dice at $n = 1000$. The per-cell baseline is selected on the same data. Against RRD and the ε-mixed Nash, fixed in advance, it is better in 15/21 cells each and worse in 6 and 4. Under Holm correction over the 21 cells, 10 and 15 gains and 5 and 2 losses stay significant (Table 16). On a shared simulation budget (Table 17) the same 12/21 lowest-mean cells and every Holm-significant sign persist; only dark hex's best solver changes, from QRE-U to exact Nash.

Table 1: Paired AUC difference of anchored QRE against the strongest baseline in each cell (mean $\pm$ the larger half-width of the paired bootstrap interval, negative = ours better; 10 seeds per cell). Cells marked $*$, †, ‡ compare against the uniform-anchored QRE, maximum-entropy Nash, and α-Rank respectively; unmarked cells compare against RRD.

domain	n=10	n=100	n=1000
matrix (30×30)	-0.19 ± 0.15	$-0.14 \pm 0.15^{*}$	$-0.15 \pm 0.11^{*}$
Goofspiel-4	$-0.06 \pm 0.03^{\dagger}$	$-0.11 \pm 0.04^{\dagger}$	$-0.12 \pm 0.03^{\dagger}$
Goofspiel-5	$-0.11 \pm 0.02^{*}$	$-0.10 \pm 0.02^{*}$	$-0.09 \pm 0.01^{*}$
Leduc poker	$+0.16 \pm 0.12^{\ddagger}$	$-0.17 \pm 0.09^{\ddagger}$	$-0.10 \pm 0.08^{*}$
liar's dice (1d4)	$-0.02 \pm 0.03^{\ddagger}$	$+0.01 \pm 0.03^{\ddagger}$	$+0.02 \pm 0.02^{\ddagger}$
Kuhn poker	$+0.04 \pm 0.05^{*}$	$+0.09 \pm 0.04^{*}$	$+0.09 \pm 0.03^{*}$
dark hex (2×2)	$+0.39 \pm 0.09^{*}$	$+0.40 \pm 0.05^{*}$	$+0.40 \pm 0.07^{*}$

The cost is speed, and it is the anchor's. On both losing domains the anchored QRE's final exploitability trails the best solver by at most 0.035 in mean: its paired terminal deficit (Table 18) excludes zero only at Kuhn $n = 100$ and vanishes on dark hex. The uniform-anchored QRE, run with the same temperature schedule, has the lowest mean AUC of the eight at every n (Tables 15–18). The

mechanism is slow adoption (Table 19): a newly appended best response receives 0.02–0.14 of the anchored meta-solution's mass against 0.44–0.53 under QRE-U, on every seed. This lower uptake fits the AUC deficit: it is anchor propagation, the quantity Corollary 6.2 holds fixed per round. The hotter-schedule deficit on Kuhn has the same source, a larger τ pulling harder toward the anchor.

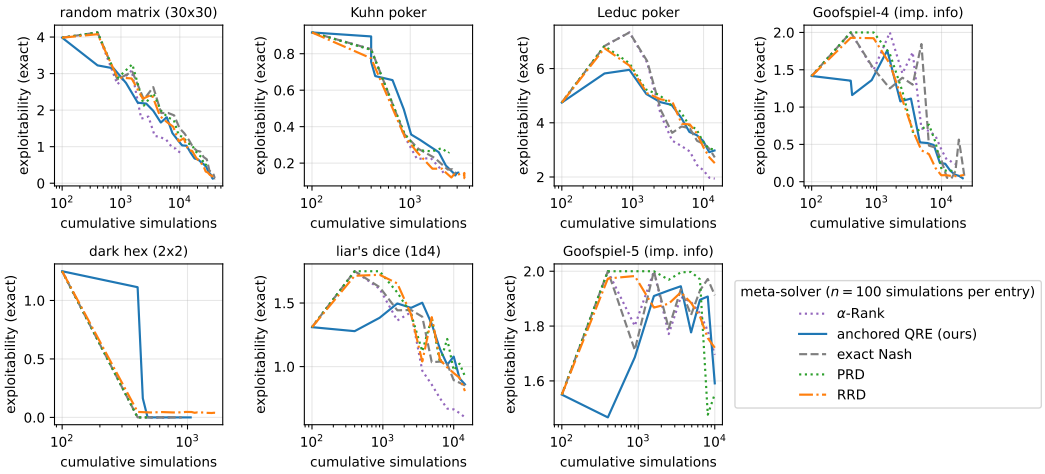


Figure 2: Exact exploitability against cumulative simulations ($n = 100$ episodes per entry, mean over seeds) for five meta-solvers on seven domains; noise enters only through the estimated meta-game.

Exact-to-noisy degradation. Rerunning all eight meta-solvers on a near-exact meta-game (exact on the matrix; 10,000 episodes per entry elsewhere; noisy-run seeds except five of ten on Leduc, Goofspiel-5 and liar's dice) keeps the wins and losses (Tables 11–12). The anchored QRE keeps the best near-exact AUC on Leduc and Goofspiel-4/5, and the uniform-anchored QRE on dark hex and Kuhn, as under noise. On the matrix, paired per seed on the round axis (Table 13), the anchored QRE ties exact Nash on the exact meta-game (-0.008; ahead on $5/10$ seeds). Under noise it leads (-0.040, -0.076, -0.020 at $n = 10, 100, 1000$; ahead on 7, 8, 6 seeds). Only exact Nash and maximum-entropy Nash end at exploitability zero, and the uniform-anchored QRE has the lowest round-axis mean throughout. QRE regularization cuts the Nash solver's output spread 2–5$\times$ on every cell (Table 14).

Deadlock, repair, and the anchor value. On a 60×60 Gaussian matrix with the anchor set to the previous round's meta-solution (10 seeds), all three branches of Theorem 5.1 are exercised (Table 3; Appendix D). With $\varepsilon = 0$ all 10 runs deadlock completely: new strategies receive negligible weight while exploitability stalls, and singleton starts freeze entirely. Any $\varepsilon > 0$ removes it, final exploitability falling as ε rises to 0.5 (Remark B.4). Pure uniform anchoring ($\varepsilon = 1$) is beaten by an interior setting (Table 3): $\varepsilon = 0.5$ is best (0.253 vs. 0.406, paired difference -0.154 ± 0.037), so the previous round's meta-solution is positive information where the meta-game is coarsest. Under fresh noise ($n \in \{50, 200\}$) the matrix and Kuhn show no significant paired difference between $\varepsilon = 0.5$ and 1.0, so the anchor's advantage concentrates on the exact meta-game (Appendix D).

Sweeping τ_0, decay, and ε at $n = 100$ against RRD (Figure 4a, Appendix E), we find the deficit grows as the schedule stays hot on Kuhn while the matrix's win persists on $9/27$; the robustness criterion *failed*.

Scaling and the terminal bound. Remark 6.3 predicts $\tau^\star \asymp \sqrt{B\kappa_n/L}$ with $L = \log(n_1/\varepsilon) + \log(n_2/\varepsilon)$, a square-root law in the meta-game noise scale $\hat{\sigma}$. The fitted scaling exponent $\beta = d \log \tau^\star / d \log \hat{\sigma}$ is 0.98 on the synthetic class ($5/6$ excluding 0.5), rejecting the predicted $1/2$; real games are flat (Figure 4b, Table 20, Appendix E). All 80 runs satisfy the concentration event and the bound (12). For the no-new rule this follows from the theorem. For the improvement-threshold rule it is an empirical envelope: that rule obeys its own guarantee $\mathrm{NashConv} < 2\theta$ (Proposition B.2) and ends with terminal NashConv 44–59% below the no-new rule (Appendix D). On Leduc with a deep Q-network oracle (Mnih et al., 2015), final-NashConv differences (0.17–0.71) lie within the oracle gap (0.53–0.75), so the ranking is unresolved, though the ε-mixed anchor keeps adding strategies throughout (Appendix D).

AI USE STATEMENT

In this work, we used generative AI tools to revise and polish the wording of the manuscript, and to write, run, and debug the experiment scripts whose results appear in the main text and appendices. We reviewed all AI-assisted work; experiment scripts produced with AI assistance were run and checked against the reported results. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

A complete description of every experiment appears in Appendix F. The pipelines behind the main-text experiments (escape study, mechanism probes, noise Pareto, hyperparameter ablation, deadlock and repair, matched degradation, and the terminal bound) are provided as an anonymized code repository in the submission bundle, from which every reported number can be regenerated on a single CPU node. We will release the full source code upon acceptance.

REFERENCES

Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *Uncertainty in Artificial Intelligence*, pp. 1–10. PMLR, 2022.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 31–80, 2024.

Kenshi Abe, Mitsuki Sakamoto, Kaito Ariu, and Atsushi Iwasaki. Asymmetric perturbation in solving bilinear saddle-point optimization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=fyae4jQ9Lw.

Ioannis Anagnostides, Ioannis Panageas, Nikolas Patris, and Tuomas Sandholm. (doubly) exponential lower bounds for follow the regularized leader in potential games. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=l6KZJO7w48.

David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. Re-evaluating evaluation. *Advances in Neural Information Processing Systems*, 31, 2018.

Ariyan Bighashdel, Yongzhao Wang, Stephen McAleer, Rahul Savani, and Frans A Oliehoek. Policy space response oracles: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7951–7961, 2024.

Branislav Bosansky, Christopher Kiekintveld, Viliam Lisy, and Michal Pechoucek. An exact double-oracle algorithm for zero-sum extensive-form games with imperfect information. *Journal of Artificial Intelligence Research*, 51:829–866, 2014.

Le Cong Dinh, Yaodong Yang, Stephen McAleer, Zheng Tian, Nicolas Perez Nieves, Oliver Slumbers, David Henry Mguni, Haitham Bou Ammar, and Jun Wang. Online double oracle. *arXiv preprint arXiv:2103.07780*, 2021.

Adam Dziwoki and Rostislav Horcik. Perturbing best responses in zero-sum games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16863–16870, 2026.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Regret circuits: Composability of regret minimizers. In *ICML*, Proceedings of Machine Learning Research, pp. 1863–1872. PMLR, 2019.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Ian Gemp, Rahul Savani, Marc Lanctot, Yoram Bachrach, Thomas Anthony, Richard Everett, Andrea Tacchetti, Tom Eccles, and János Kramár. Sample-based approximation of nash in large many-player games via gradient descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 507–515, 2022.

Josef Hofbauer and William H Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning*, pp. 9695–9728. PMLR, 2022.

Christopher Kiekintveld and Michael P Wellman. Selecting strategies using empirical game models: an experimental analysis of meta-strategies. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pp. 1095–1101, 2008.

Christine Konicki, Mithun Chakraborty, and Michael P Wellman. Exploiting extensive-form structure in empirical game-theoretic analysis. In *International Conference on Web and Internet Economics*, pp. 132–149. Springer, 2022.

Harold W. Kuhn. A simplified two-person poker. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume I*, volume 24 of *Annals of Mathematics Studies*, pp. 97–103. Princeton University Press, 1950.

Marc Lanctot, Vinícius Flores Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *NIPS*, pp. 4190–4203, 2017.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Boning Li and Longbo Huang. Agents that certify their own exploits: Confidence-scheduled restricted responses for safe opponent exploitation. *arXiv preprint arXiv:2607.28520*, 2026a.

Boning Li and Longbo Huang. Real-time parallel counterfactual regret minimization. *arXiv preprint arXiv:2605.19928*, 2026b.

Boning Li, Yu Chen, and Longbo Huang. Av-aivat: 74x cheaper agent evaluation with certified anytime-valid stopping in imperfect-information games. *arXiv preprint arXiv:2608.06362*, 2026a.

Boning Li, Yu Chen, and Longbo Huang. Correlated chance sampling for monte carlo counterfactual regret minimization. *arXiv preprint arXiv:2607.27035*, 2026b.

Wataru Masaka, Mitsuki Sakamoto, Kenshi Abe, Kaito Ariu, Tuomas Sandholm, and Atsushi Iwasaki. On the power of perturbation under sampling in solving extensive-form games. *arXiv preprint arXiv:2501.16600*, 2025.

Stephen McAleer, John B Lanier, Roy Fox, and Pierre Baldi. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games. *Advances in neural information processing systems*, 33:20238–20248, 2020.

Stephen McAleer, John B Lanier, Kevin A Wang, Pierre Baldi, and Roy Fox. Xdo: A double oracle algorithm for extensive-form games. *Advances in Neural Information Processing Systems*, 34: 23128–23139, 2021.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1(1):9–41, 1998.

Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 525–533, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

H Brendan McMahan, Geoffrey J Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 536–543, 2003.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. URL https://EconPapers.repec.org/RePEc:nat:nature:v:518:y:2015:i:7540:d:10.1038_nature14236.

Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Pérolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Bkl5kxrKDr.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific reports*, 9(1):9937, 2019.

Ioannis Panageas, Nikolas Patris, Stratis Skoulakis, and Volkan Cevher. Exponential lower bounds for fictitious play in potential games. *Advances in Neural Information Processing Systems*, 36: 32401–32423, 2023.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022. doi: 10.1126/science.add4679. URL https://www.science.org/doi/abs/10.1126/science.add4679.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Sheldon M Ross. Goofspiel—the game of pure strategy. *Journal of Applied Probability*, 8(3): 621–625, 1971.

Max Olan Smith and Michael P Wellman. Co-learning empirical games and world models. *arXiv preprint arXiv:2305.14223*, 2023.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023a.

Samuel Sokota, Ryan D'Orazio, Chun Kai Ling, David J Wu, J Zico Kolter, and Noam Brown. Abstracting imperfect information away from two-player zero-sum games. In *International Conference on Machine Learning*, pp. 32169–32193. PMLR, 2023b.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 550–558, 2005.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

Xiaohang Tang, Le Cong Dinh, Stephen Marcus McAleer, and Yaodong Yang. Regret-minimizing double oracle for extensive-form games. In *International Conference on Machine Learning*, pp. 33599–33615. PMLR, 2023.

Xiaohang Tang, Chiyuan Wang, Chengdong Ma, Ilija Bogunovic, Stephen McAleer, and Yaodong Yang. Sample-efficient regret-minimizing double oracle in extensive-form games. *CoRR*, abs/2411.00954, 2024.

Yongzhao Wang and Michael P Wellman. Regularization for strategy exploration in empirical game-theoretic analysis. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2484–2486, 2023.

Michael P Wellman. Methods for empirical game-theoretic analysis. In *AAAI*, pp. 1552–1556, 2006.

Mason Wright, Yongzhao Wang, and Michael P Wellman. Iterated deep reinforcement learning in games: History-aware training for improved stability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 617–636, 2019.

Brian Hu Zhang and Tuomas Sandholm. Exponential lower bounds on the double oracle algorithm in zero-sum games. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3032–3039, 2024.

Junyu Zhang, Feihong Yang, Jian Wang, Chao Wang, and Xudong Zhang. Global policy-space response oracles for two-player zero-sum games. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=DIytYt7tdI`.

Ming Zhou, Jingxiao Chen, Ying Wen, Weinan Zhang, Yaodong Yang, Yong Yu, and Jun Wang. Efficient policy space response oracles. *arXiv preprint arXiv:2202.00633*, 2022.

A Proofs for Section 4

Definition A.1 (Interior MSS)**.** An MSS is *interior* if it always returns (μ^t, ν^t) with full support on the current populations. By (3), the anchored QRE with interior anchors is interior for every temperature $\tau > 0$.

Lemma A.2 (Forcing)**.** *Under any interior MSS, at every round, for every attack-state bit $j \in \{2, \dots, k\}$: (i) if bit j is attacked, every exact best response of player 1 has $x_j = 1$; (ii) if bit j is dirty, every exact best response of player 2 has $y_j = 1$, so j is attacked from the next round on; (iii) dirty and attacked are absorbing. Player 1 is tied exactly on the unattacked attack bits and player 2 exactly on the non-dirty attack bits; both players are tied simultaneously exactly on the virgin bits (neither dirty nor attacked). Bit-1 ties occur only at exact marginal balance.*

Lemma A.3 (QRE mass lower bound)**.** *For a zero-sum matrix game A with $|A_{ij}| \le B$, the anchored QRE $\mathrm{QRE}_{\tau,\rho}$ assigns, to every pure strategy s with $\rho_s > 0$ (so, for an interior anchor, to every current-population strategy), mass at least*

$$\rho_s \, e^{-2B/\tau}. \tag{13}$$

Proof. The logit fixed point reads $\mu_i = \rho_i \, e^{(A\nu)_i/\tau}/Z$ with $Z = \sum_{i'} \rho_{i'} \, e^{(A\nu)_{i'}/\tau}$; since $(A\nu)_i \in [-B, B]$, the numerator is at least $\rho_i e^{-B/\tau}$ and $Z \le e^{B/\tau}$, so $\mu_i \ge \rho_i e^{-2B/\tau}$. The column player is symmetric on $-A^\top$. □ □

Proof of Lemma 4.2 (marginal form and gap identity). Since μ and ν are independent and u is a sum of per-bit terms,

$$\mathbb{E}\,[m(x_1, y_1)] = \Pr[x_1{=}y_1] - \Pr[x_1{\neq}y_1] = (2p_1 - 1)(2q_1 - 1),$$
$$\mathbb{E}\,[g(x_j, y_j)] = 1 - 2\Pr[x_j{=}0]\Pr[y_j{=}1],$$

which gives the stated form of $u(\mu, \nu)$. For the best-response values,

$$\max_x u(x, \nu) = \frac{1}{k}\Big[\max\{2q_1 - 1,\, 1 - 2q_1\} + \sum_{j\ge 2} \max\{1,\, 1 - 2q_j\}\Big] = \frac{1}{k}\Big[|2q_1 - 1| + (k-1)\Big],$$

using $q_j \ge 0$, and

$$\min_y u(\mu, y) = \frac{1}{k}\Big[-|2p_1 - 1| + \sum_{j\ge 2} \min\{1,\, 2p_j - 1\}\Big] = \frac{1}{k}\Big[-|2p_1 - 1| + \sum_{j\ge 2}(2p_j - 1)\Big],$$

using $p_j \le 1$. Subtracting $u(\mu,\nu)$ from the first display and simplifying gives

$$\mathrm{gap}_1(\mu,\nu) = \frac{1}{k}\Big[\,|2q_1-1| - (2p_1-1)(2q_1-1) + 2\sum_{j\ge 2} q_j(1-p_j)\Big], \tag{14}$$

$$\mathrm{gap}_2(\mu,\nu) = \frac{1}{k}\Big[\,|2p_1-1| + (2p_1-1)(2q_1-1) + 2\sum_{j\ge 2}(1-p_j)(1-q_j)\Big], \tag{15}$$

where the second line uses $1-2q_j(1-p_j)-(2p_j-1) = 2(1-p_j)(1-q_j)$. Adding the two, the terms $\mp(2p_1-1)(2q_1-1)$ cancel and $2q_j(1-p_j)+2(1-p_j)(1-q_j) = 2(1-p_j)$, giving (6).

The claims of Definition 4.1 follow: $\mathrm{NashConv} = 0$ forces $p_1 = q_1 = \frac{1}{2}$ and $p_j = 1$ for $j \ge 2$, with q_j unconstrained; plugging the equilibrium marginals into u gives $v^\star = (k-1)/k$. □

Lemma A.4 (Best-response structure)**.** *With BR_i the pure best-response sets of Section 3, against ν with marginals q the set $\mathrm{BR}_1(\nu)$ is the product over bits of per-bit argmax sets: bit 1 is $\mathbb{I}[q_1 > \frac{1}{2}]$ (both actions optimal iff $q_1 = \frac{1}{2}$); bit $j \ge 2$ is 1 if $q_j > 0$ and is a tie iff $q_j = 0$. Symmetrically, for $\mathrm{BR}_2(\mu)$: bit 1 is $\mathbb{I}[p_1 < \frac{1}{2}]$ (tie iff $p_1 = \frac{1}{2}$); bit $j \ge 2$ is 1 if $p_j < 1$ and is a tie iff $p_j = 1$.*

$u(x,\nu) = \frac{1}{k}\sum_j \phi_j(x_j;q)$ is separable across bits, so $\arg\max_x u(x,\nu)$ is the product of the per-bit argmax sets. The per-bit comparisons are: bit 1, $\phi_1(1)-\phi_1(0) = (2q_1-1)-(1-2q_1) = 2(2q_1-1)$; bit $j \ge 2$, $\phi_j(1)-\phi_j(0) = 1-(1-2q_j) = 2q_j$. Player 2 is symmetric with payoff $-u$: bit 1 difference $-2(2p_1-1)$ favoring $y_1 = 1$ iff $p_1 < \frac{1}{2}$; bit $j \ge 2$ difference $2(1-p_j)$ favoring $y_j = 1$ iff $p_j < 1$. □

Proof of Lemma A.2 (forcing). Fix an attack-state bit $j \ge 2$. If bit j is attacked at round t, some $y \in \Pi_2^t$ has $y_j = 1$; interiority gives $\nu_y^t > 0$, hence $q_j^t \ge \nu_y^t > 0$, and by Lemma A.4 every exact best response of player 1 has $x_j = 1$ (the per-bit gain $2q_j^t/k$ is strictly positive). Claim (ii) is symmetric: dirtiness gives $p_j^t \le 1 - \mu_x^t < 1$ for the zero-carrying x, forcing $y_j = 1$ in every best response of player 2; the forced response either is appended at the end of round t or already lies in Π_2^t, and in both cases bit j is attacked at round $t+1$. Claim (iii) holds because populations only grow. The separate bit-1 statement is the exact marginal condition of Lemma A.4 ($q_1 = \frac{1}{2}$, respectively $p_1 = \frac{1}{2}$), where attack-state forcing does not apply. □

Proof of Theorem 4.3 (polynomial escape). Throughout, rounds are indexed $t = 0, 1, \ldots$ from the initial populations (Π_1^0, Π_2^0) (Algorithm 1 numbers the same rounds from 1), and "round t" refers to the state at the beginning of the round: populations (Π_1^t, Π_2^t), meta-solution (μ^t, ν^t) with marginals (p^t, q^t), best responses computed against (μ^t, ν^t) and appended (if new) at the end of the round. Recall $e_1 := 1^k \in \Pi_1^t$ for all t and write $e_0 := (0, 1, \ldots, 1)$.

Achievers. Say $x \in \Pi_1^t$ is a 1-*achiever* at round t if $u(x,\nu^t) = \max_{x' \in \{0,1\}^k} u(x',\nu^t)$. By Lemma A.4, x is a 1-achiever iff (a) x_1 is bit-1 optimal (automatic when $q_1^t = \frac{1}{2}$), and (b) for every $j \ge 2$, either $x_j = 1$ or $q_j^t = 0$. By interiority, $q_j^t = 0$ iff bit j is unattacked. 2-*achievers* $y \in \Pi_2^t$ are defined symmetrically: y_1 bit-1 optimal (automatic when $p_1^t = \frac{1}{2}$) and, for every $j \ge 2$, either $y_j = 1$ or $p_j^t = 1$; by interiority, $p_j^t = 1$ iff bit j is not dirty.

Step 1: an achiever caps the gap below the threshold. Suppose a 1-achiever exists at round t. Then the full and restricted best-response values against ν^t coincide, so, writing $\rho = \rho_1^t$ for the player-1 anchor and using that μ^t maximizes $\mu \mapsto \mu^\top A^t \nu^t - \tau_t \mathrm{KL}(\mu\|\rho)$, comparison with the vertex at any restricted best response $x^\star$ gives

$$\mathrm{gap}_1(\mu^t,\nu^t) = u(x^\star,\nu^t) - u(\mu^t,\nu^t) \le \tau_t\big(\mathrm{KL}(\delta_{x^\star}\|\rho) - \mathrm{KL}(\mu^t\|\rho)\big) \le \tau_t \log\tfrac{1}{\rho_{\min}^t} \overset{(7)}{\le} \tfrac{1}{k} < \tfrac{2}{k}.$$

The same bound holds for gap_2 when a 2-achiever exists. Hence, at a *blocked* round ($\max_i \mathrm{gap}_i \ge 2/k$), at least one of the two achiever sets is empty. A non-blocked round triggers termination, so every round before termination is blocked, and it suffices to bound the number of blocked rounds by $4k-1$.

Step 2: 1-blocked rounds. Call round t 1-*blocked* if no 1-achiever exists. Since $e_1 \in \Pi_1^t$ satisfies (b) trivially, this forces $q_1^t < \frac{1}{2}$ strictly. The best response x^t is then forced to $x_1^t = 0$ and $x_j^t = 1$ on

all attacked bits; its zero set $\mathcal{Z}_t := \{j \geq 2 : x_j^t = 0\}$ is contained in the unattacked bits. Moreover $x^t \notin \Pi_1^t$: any element of Π_1^t equal to x^t would satisfy (a) ($x_1 = 0$ is optimal at $q_1^t < \frac{1}{2}$) and (b) (1 on attacked bits, arbitrary on unattacked), contradicting 1-blockedness. So x^t is appended.

Subcase (a): $\mathcal{Z}_t = \emptyset$, i.e. $x^t = e_0$. From the next round on, $\{e_0, e_1\} \subseteq \Pi_1$, and one of them satisfies (a) at every round while both satisfy (b) (all-ones tails); hence no future round is 1-blocked. This subcase occurs at most once.

Subcase (b): $\mathcal{Z}_t \neq \emptyset$. Unattacked bits split into *virgin* bits (neither dirty nor attacked) and dirty-but-unattacked bits. By Lemma A.2(ii), every bit dirty at round $t-1$ is attacked at round t; hence any dirty-but-unattacked bit at round t became dirty exactly at the end of round $t-1$, i.e. is a zero placed by x^{t-1} on a bit that was virgin at round $t-1$ (zeros are only placeable on unattacked bits, and a previously dirty bit would already be dirty). Therefore, at a 1-blocked round of subcase (b), either

(V) $\mathcal{Z}_t$ contains a virgin bit, which becomes dirty, and dirtiness is absorbing, so this happens at most $k-1$ times over the whole run; or

(R) $\mathcal{Z}_t$ consists entirely of bits that x^{t-1} zeroed while virgin at round $t-1$; we charge round t to round $t-1$, which is a round whose player-1 addition consumed at least one virgin bit. Each such round is charged at most once (by its successor), and there are at most $k-1$ virgin-consuming rounds in total.

Hence the number of 1-blocked rounds is at most $(k-1) + (k-1) + 1 = 2k-1$.

Step 3: 2-blocked rounds. Call round t *2-blocked* if no 2-achiever exists. The best response y^t is forced to $y_j^t = 1$ on every dirty bit (Lemma A.2), and, when $p_1^t \neq \frac{1}{2}$, to the bit-1 optimal y_1^t; as in Step 2, $y^t \notin \Pi_2^t$ (a copy in the population would be a 2-achiever), so y^t is appended, and y^t is a 2-achiever *certificate* for the configuration $c(t) := (y_1^t, \mathrm{dirty}(t))$: at any future round s with $\mathrm{dirty}(s) = \mathrm{dirty}(t)$ and bit-1 requirement compatible with y_1^t (i.e. $p_1^s = \frac{1}{2}$ or $\mathbb{I}[p_1^s < \frac{1}{2}] = y_1^t$), y^t is a 2-achiever: dirty bits carry 1s, non-dirty bits satisfy $p_j^s = 1$, and bit 1 matches. Hence each configuration can be 2-blocked at most once. The dirty sets form an increasing chain with at most k distinct values ($\emptyset$ plus at most $k-1$ strict growths), and there are two bit-1 requirements, so the number of 2-blocked rounds is at most $2k$.

Step 4: conclusion. Every pre-termination round is 1-blocked or 2-blocked, so there are at most $(2k-1) + 2k = 4k-1$ of them; the algorithm therefore terminates with $\max_i \mathrm{gap}_i < 2/k$ by round $4k$. □

Remark A.5. The proof uses interiority only through the equivalences "$q_j = 0$ iff unattacked" and "$p_j = 1$ iff not dirty", i.e. the meta-solver may be *any* rule whose output has full support, provided its restricted gaps obey the Step-1 bias bound. The anchored QRE supplies both properties simultaneously for every τ_t satisfying (7).

Lemma A.6 (Quantitative forcing)**.** *Let $B := \max_{x,y} |u(x,y)|$ and write $m_{\min}^t := \rho_{\min}^t e^{-2B/\tau_t}$, with $\rho_{\min}^t$ the minimum anchor mass over both current populations. Under any interior MSS whose per-strategy mass is at least $m_{\min}^t$ (guaranteed for the anchored QRE by Lemma A.3), every* pure *η-approximate best response, with value loss at most η measured against the true game payoff and with $\eta < 2m_{\min}^t/k$, satisfies: (i) for every attack-state bit $j \geq 2$ that is attacked, $x_j = 1$; (ii) for every attack-state bit $j \geq 2$ that is dirty, $y_j = 1$.*

Proof. Fix an attack-state bit $j \geq 2$. If bit j is attacked, some $y^\star \in \Pi_2^t$ has $y_j^\star = 1$, so $q_j^t \geq \nu_{y^\star} \geq m_{\min}^t$ and, by Lemma A.4, the per-bit gain of setting $x_j = 1$ is $2q_j^t/k \geq 2m_{\min}^t/k > \eta$; a *pure* response with $x_j = 0$ would lose more than the η budget against the true payoff, so it cannot be an η-approximate best response. Symmetrically, if bit j is dirty, some $x^\star \in \Pi_1^t$ has $x_j^\star = 0$, so $p_j^t \leq 1 - \mu_{x^\star} \leq 1 - m_{\min}^t$ and the per-bit gain of $y_j = 1$ for player 2 is $2(1-p_j^t)/k \geq 2m_{\min}^t/k > \eta$. □ □

Theorem A.7 (η-robust escape under a novelty-constrained oracle)**.** *Run DO on G_k from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$ with MSS $= \mathrm{QRE}_{\tau_t,\rho^t}$ for any interior anchors ρ^t (full support on the current populations; the initial point mass is the relative-interior point of its singleton simplex) and any positive temperature schedule $\tau_t > 0$ satisfying* (7)*. At every round and for each player $i \in \{1,2\}$ the oracle returns a* pure *η-approximate best response, with value loss at most η per best-response*

computation measured against the true payoff, under arbitrary adversarial tie-breaking otherwise. Player 2's response is additionally required to play an exactly optimal bit 1 ($y_1^t = 1$ iff $p_1^t < \frac{1}{2}$; at $p_1^t = \frac{1}{2}$ bit 1 is unconstrained). Both responses obey the novelty constraint *of the double-oracle append rule, imposed per player on that player's admissible class: letting $\mathcal{C}_2^t$ be the pure η-approximate best responses of player 2 with exactly optimal bit 1 and $\mathcal{C}_1^t$ the pure η-approximate best responses of player 1, whenever $\mathcal{C}_i^t$ contains a strategy outside the current population Π_i^t, the player-i response returned is one that does. On G_k every payoff satisfies $|u| \le 1$, so $B = 1$, and we assume*

$$\eta \;<\; \frac{2\,\rho_{\min}^t\, e^{-2/\tau_t}}{k} \tag{16}$$

for every round t before termination. Then DO reaches $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ within at most $4k$ rounds.

Proof. We rerun the proof of Theorem 4.3 with Lemma A.6 in place of Lemma A.2 and with every best response supplied by the η-approximate oracle; write $m_{\min}^t := \rho_{\min}^t e^{-2/\tau_t}$ (the bound $B = 1$ on G_k). Every conclusion below holds for whatever pure response the adversarial oracle returns.

Step 1 (achiever caps the gap; unchanged). If an *exact* best response of player i lies in Π_i^t, comparison in the QRE objective bounds $\mathrm{gap}_i(\mu^t, \nu^t) \le \tau_t \log(1/\rho_{\min}^t) \le 1/k < 2/k$ exactly as in the proof of Theorem 4.3; this bound uses only the KL structure of the meta-solver, so it is unaffected by the oracle approximation. Hence at a *blocked* round ($\max_i \mathrm{gap}_i \ge 2/k$) the two players cannot both have an in-population exact best response, i.e. at least one player is blocked; every pre-termination round is blocked.

Step 2 (1-blocked rounds). Call round t 1-*blocked* if no exact player-1 best response lies in Π_1^t. Player 1's exact best response is then population-external and in particular an η-approximate one, hence a member of $\mathcal{C}_1^t$; the novelty constraint on $\mathcal{C}_1^t$ therefore forces the oracle to return a population-external pure $x^t \in \mathcal{C}_1^t$. On every attacked bit $j \ge 2$, Lemma A.6(i) forces $x_j^t = 1$ (setting $x_j^t = 0$ costs $2q_j^t/k \ge 2m_{\min}^t/k > \eta$), so the zero set $\mathcal{Z}_t := \{j \ge 2 : x_j^t = 0\}$ is contained in the unattacked bits. Since $e_1 \in \Pi_1^t$ and $x^t \notin \Pi_1^t$, $x^t \ne e_1$, so either $\mathcal{Z}_t = \emptyset$ (then $x^t = e_0$) or $\mathcal{Z}_t \ne \emptyset$.

Subcase (a): $\mathcal{Z}_t = \emptyset$, i.e. $x^t = e_0$, which is appended. Once $\{e_0, e_1\} \subseteq \Pi_1$, at every later round one of them is an exact player-1 best response (e_0 has the bit-1 optimum $x_1 = 0$ when $q_1 < \frac{1}{2}$, e_1 has $x_1 = 1$ when $q_1 \ge \frac{1}{2}$; both have all-ones tails), so no future round is 1-blocked. This subcase occurs at most once.

Subcase (b): $\mathcal{Z}_t \ne \emptyset$. Unattacked bits split into *virgin* bits and dirty-but-unattacked bits. By Lemma A.6(ii), every bit dirty at round $t-1$ is attacked at round t (the player-2 response y^{t-1} carries a 1 on it and is appended or already present), so any dirty-but-unattacked bit at round t became dirty exactly at the end of round $t-1$, i.e. is a zero placed by x^{t-1} on a bit that was virgin at round $t-1$ (zeros are placeable only on unattacked bits, and an earlier-dirty bit would already be attacked). Therefore, at a 1-blocked round of subcase (b), either

(V) $\mathcal{Z}_t$ contains a virgin bit, which becomes dirty; dirtiness is absorbing, so at most $k-1$ such rounds occur over the run; or

(R) $\mathcal{Z}_t$ consists entirely of bits that x^{t-1} zeroed while virgin at round $t-1$; we charge round t to round $t-1$, a virgin-consuming round charged by its successor at most once, and there are at most $k-1$ virgin-consuming rounds in total.

The number of 1-blocked rounds is at most $(k-1) + (k-1) + 1 = 2k-1$.

Step 3 (2-blocked rounds). Call round t 2-*blocked* if no exact player-2 best response lies in Π_2^t. Every exact player-2 best response is then population-external and, being exact, has an exactly optimal bit 1 and value loss $0 \le \eta$, hence is a member of $\mathcal{C}_2^t$; the novelty constraint on $\mathcal{C}_2^t$ therefore forces the oracle to return a population-external pure $y^t \in \mathcal{C}_2^t$. By Lemma A.6(ii) every dirty bit carries $y_j^t = 1$. By the bit-1 exactness condition of the statement, y_1^t is always an exact bit-1 optimum, so the certificate records a bit-1 value compatible with the round-t marginal. Record the certificate y^t for the configuration $c(t) := (y_1^t, \mathrm{dirty}(t))$: at any future round s with $\mathrm{dirty}(s) = \mathrm{dirty}(t)$ and bit-1 requirement compatible with y_1^t (i.e. $p_1^s = \frac{1}{2}$ or $\mathbb{I}[p_1^s < \frac{1}{2}] = y_1^t$), y^t is an *exact* player-2 best

response: dirty bits carry 1s, non-dirty bits have $p_j^s = 1$ (interiority), and bit 1 matches. This property depends only on the structure of y^t (forcing on dirty bits), not on any optimality of y^t at round t, and y^t remains in the population permanently. Now if two 2-blocked rounds $t < s$ had $c(t) = c(s)$, then $y_1^t = y_1^s$, and the exactness condition makes y_1^s the exact bit-1 optimum at round s, so $p_1^s = \frac{1}{2}$ or $\mathbb{I}[p_1^s < \frac{1}{2}] = y_1^s = y_1^t$: the bit-1 requirement is compatible, y^t is an exact player-2 best response at round s (still in the population), contradicting that s is 2-blocked. Hence each configuration is 2-blocked at most once; the dirty sets form an increasing chain with at most k distinct values ($\emptyset$ plus at most $k-1$ strict growths) and bit 1 takes two values, so at most $2k$ rounds are 2-blocked.

Step 4 (conclusion). Every pre-termination round is 1-blocked or 2-blocked, so at most $(2k-1) + 2k = 4k - 1$ rounds precede termination; DO reaches $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ within at most $4k$ rounds. □ □

Remark A.8 (Unrestricted approximate oracles fail at large error). The proof uses the novelty constraint. For an unrestricted η-approximate oracle, which may return a strategy already in the population, the conclusion fails whenever $\eta \geq 2/k$. At the initial populations the meta-solution is the pure profile $(1^k, 0^k)$ with $\mathrm{gap}_1 = 2/k$ and $\mathrm{gap}_2 = 0$ (Lemma 4.2): every exact best response of player 1 plays $x_1 = 0$ and attains value 1, while 1^k attains $1 - 2/k$, a shortfall of exactly $2/k$; 0^k is exact for player 2. So, under the strict termination rule of Theorem 4.3 (a round at exactly $2/k$ does not stop the run), each incumbent strategy is an η-approximate best response. An adversarial oracle may return the incumbents forever: nothing is ever appended, and the populations and the meta-solution never change, so the run never reaches $\max_i \mathrm{gap}_i < 2/k$. The probe of Table 10 realizes exactly this failure at $\eta \in \{1, 10\}$ ($0/3$ seeds converge on every k).

Corollary A.9 (ε-mixed Nash: escape with inverse-polynomial tolerance)**.** *Run DO on G_k from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$ with the meta-solver*

$$\mu^t = (1-\varepsilon)\,\mu_t^\star + \varepsilon\,\mathrm{unif}(\Pi_1^t), \qquad \nu^t = (1-\varepsilon)\,\nu_t^\star + \varepsilon\,\mathrm{unif}(\Pi_2^t), \tag{17}$$

where $(\mu_t^\star, \nu_t^\star)$ is any *Nash equilibrium of the restricted game A^t, adversarially selected if desired, and $0 < \varepsilon \leq 1/(4k)$. (i) With exact best responses under arbitrary adversarial tie-breaking, DO reaches $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ within at most $4k$ rounds. (ii) The same holds under the novelty-constrained pure η-approximate oracle of Theorem A.7 whenever $\eta < 2\varepsilon/(k \max_i |\Pi_i^t|)$ at every pre-termination round; since $|\Pi_i^t| \leq t + 1 \leq 4k$ there, $\eta < \varepsilon/(2k^2)$ suffices, and for $\varepsilon = 1/(4k)$ any $\eta < 1/(8k^3)$ does.*

Proof. Every strategy of Π_i^t receives mass at least $\varepsilon/|\Pi_i^t| \geq m^t := \varepsilon/\max_i |\Pi_i^t| > 0$, so the meta-solver is interior (Definition A.1) with per-strategy floor m^t. By the preceding remark it remains to re-derive Step 1, the only step that uses the QRE objective. Suppose a 1-achiever exists at round t, so the full and restricted best-response values against ν^t coincide, and let v_t be the value of A^t. Since $\mu_t^\star$ is maximin on $\Pi_1^t \times \Pi_2^t$, $u(\mu_t^\star, y) \geq v_t$ for every $y \in \Pi_2^t$, hence $u(\mu_t^\star, \nu^t) \geq v_t$, and with $|u| \leq 1$,

$$u(\mu^t, \nu^t) = (1-\varepsilon)\,u(\mu_t^\star, \nu^t) + \varepsilon\,u(\mathrm{unif}(\Pi_1^t), \nu^t) \;\geq\; (1-\varepsilon)\,v_t - \varepsilon \;\geq\; v_t - 2\varepsilon.$$

For every $x \in \Pi_1^t$, $u(x, \nu_t^\star) \leq v_t$ (minimax), so $u(x, \nu^t) = (1-\varepsilon)\,u(x, \nu_t^\star) + \varepsilon\,u(x, \mathrm{unif}(\Pi_2^t)) \leq (1-\varepsilon)\,v_t + \varepsilon \leq v_t + 2\varepsilon$. Hence $\mathrm{gap}_1(\mu^t, \nu^t) \leq 4\varepsilon \leq 1/k < 2/k$, and symmetrically for gap_2 when a 2-achiever exists. Steps 2–4 use interiority only through Lemma A.2 (exact case) or Lemma A.6 with floor m^t (approximate case, which needs $\eta < 2m^t/k$), and the certificate accounting is unchanged. At most $4k - 1$ rounds precede termination, each appending at most one strategy per player, so $|\Pi_i^t| \leq t + 1 \leq 4k$ at every pre-termination round. □ □

Remark A.10 (Precision and the QRE mass floor). Under (7), $e^{-2/\tau_t} \leq (\rho_{\min}^t)^{2k}$, so the QRE floor $\rho_{\min}^t e^{-2/\tau_t}$ of Lemma A.3 and the tolerance (16) of Theorem A.7 are exponentially small in k (Appendix E evaluates the tolerance along the probe schedule). Exact QRE computation is assumed. Corollary A.9 locates the exponent in the QRE's mass floor: the charging argument needs only a floor and a restricted gap at most $1/k$, and the ε-mixed Nash meets both with inverse-polynomial constants and a rational, LP-computable solution. Probe P1 of Section 7 runs this rule with $\varepsilon = 10^{-2} \leq 1/(4k)$ for every tested $k \leq 14$ and escapes in three rounds. The anchored QRE contributes the last-iterate anchor and the Lipschitz stability of Section 6 on top of the escape: on the 1×2 example of Remark C.2 the ε-mixed Nash still jumps by ℓ_1-diameter $2(1-\varepsilon)$.

Remark A.11 (Where the 2^{k-1} went). The adversarial schedule of Proposition A.14 relies on *zero-mass blindness*: a tie on an attack bit j survives only while no strategy with positive meta-solver mass attacks j (by Lemma A.4 the player-1 tie at bit j is exactly the condition $q_j = 0$, coordinate by coordinate). An interior MSS makes each tie structural: player 1 stays tied on attack bit j exactly while $q_j = 0$ (no positive-mass strategy attacks j) and player 2 stays tied exactly while $p_j = 1$ (no positive-mass strategy zeros j); each certificate, once added, is permanent. Certificates live in a budget linear in k, giving at most $4k - 1$ blocked rounds: the run is $O(k)$ *independently of how ties are broken*.

Remark A.12 (Exactness assumption). Theorem 4.3 assumes exact best responses and exact QRE computation: the forcing lemma treats any positive mass as decisive, so an approximate oracle that ignores payoff advantages of order q_j/k can break it. Theorem A.7 quantifies the admissible error, exponentially small for the QRE under (7) and inverse-polynomial for the ε-mixed Nash (Corollary A.9, Remark A.10). The theorem's scope is exact tabular DO; Appendix D probes the approximate case empirically and finds it negative. Estimation noise in the meta-game is a different failure mode and does not reintroduce the worst case: with an exact oracle and n fresh simulations per entry, the $4k$ escape survives with probability $1 - \delta$ (Corollary C.3).

Remark A.13 (Tightness and the delaying adversary). The adversary's only lever is spending virgin bits, and three implemented delaying oracles, including one that directly minimizes virgin-bit consumption (the accounting variable of the proof), never exceed 4 rounds for $k \leq 14$, with a constant fitting the observed rounds better than $\sqrt{k}$ or linear growth (Section 7): the interior meta-solution leaves the oracle almost no ties to poison. Whether the constant 4 in the $4k$ bound is optimal, or whether interiority in fact pins escape to $O(1)$ rounds under adversarial tie-breaking (as every implemented oracle suggests), remains open. The proof gives the $4k - 1$ upper bound.

Proposition A.14 (Exponential lower bound; Zhang & Sandholm 2024)**.** *For $k \geq 2$, started from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$, there is an adversarial selection of restricted Nash equilibria and of best-response ties under which DO sustains $\max_i \mathrm{gap}_i = 2/k$ for $2^{k-1} - 1$ consecutive rounds on G_k.*

Remark A.15 (Boundary of the mechanism: POSGs). The POSG construction of Zhang & Sandholm (2024) admits a *unique* restricted equilibrium and a *unique* best response at every round. Uniqueness constrains Nash meta-solvers: DO cannot escape the construction under any selection of restricted equilibria or best-response ties. An arbitrary non-Nash MSS such as a QRE faces no such constraint, since its trajectory need not follow the unique restricted equilibrium. Perturbation rescues DO where the hardness is selection-driven.

B Proofs for Section 5

Proof of Theorem 5.1 (anchoring deadlock). $\mathrm{KL}(\mu\|\rho_1^t) = +\infty$ unless $\mathrm{supp}(\mu) \subseteq \mathrm{supp}(\rho_1^t)$, so the saddle point of (2) coincides with the anchored QRE of the sub-meta-game restricted to $\mathrm{supp}(\rho^t)$. Newly appended strategies receive zero anchor mass by construction, so $\mathrm{supp}(\rho^t) = \mathrm{supp}(z^{t-1})$, and by induction $\mathrm{supp}(z^t) \subseteq \mathrm{supp}(z^0)$ for all t; the exploitability lower bound follows. Within the trapped support the anchor still moves: $\rho^t = z^{t-1}$ implements an implicit entropic (multiplicative) proximal update by (3), so the profile itself need not be constant in general. With singleton initial supports, however, the only distribution on $\mathrm{supp}(z^0)$ is the point mass, so $z^t = z^0$ for every t and the trajectory is frozen. The argument tracks only the supports of the anchor recursion, never the payoff matrix the solver faces, so it holds identically for estimated meta-games (the statement is not restricted to exact ones). □

Proof of Lemma 5.3 (mass floor). By (3) applied to the matrix $\hat{A}^t$ (each entry an average of n simulations from $[-B, B]$, hence in $[-B, B]$), the meta-solution satisfies $\mu_x = \rho_{1,x} e^{(\hat{A}\nu)_x/\tau} / \sum_{x'} \rho_{1,x'} e^{(\hat{A}\nu)_{x'}/\tau} \geq \rho_{1,x}\, e^{((\hat{A}\nu)_x - \max_{x'}(\hat{A}\nu)_{x'})/\tau} \geq (\varepsilon/n_1)\, e^{-2B/\tau}$, using $\rho_{1,x} \geq \varepsilon/n_1$; likewise for ν. The bound depends only on the anchor floor and on the entry range $[-B, B]$, so it holds identically for exact meta-games. □

Proof of Proposition 5.4 (progress and termination). If player 1's exact best response $x^\star$ is already in Π_1^t, then the restricted and full best-response values against ν^t coincide. Since μ^t maximizes

$\mu^\top A^t \nu^t - \tau_t \mathrm{KL}(\mu \| \rho_1^t)$, comparing with the vertex $\delta_{x^\star}$ gives

$$\begin{aligned}\mathrm{gap}_1(\mu^t, \nu^t) = u(x^\star, \nu^t) - u(\mu^t, \nu^t) \;&\le\; \tau_t\big(\mathrm{KL}(\delta_{x^\star} \| \rho_1^t) - \mathrm{KL}(\mu^t \| \rho_1^t)\big) \\ &\le\; \tau_t \log \tfrac{1}{\rho_{1,x^\star}^t} \;\le\; \tau_t \log \tfrac{n_1}{\varepsilon},\end{aligned}$$

and symmetrically for player 2. If both best responses are already present, both gaps obey (9) and the stated stopping rule fires; otherwise a new strategy is appended, which can happen at most $|S_1| + |S_2|$ times. □

Corollary B.1 (Annealing reaches any prescribed precision)**.** *Under the hypotheses of Proposition 5.4 (exact meta-games, $\varepsilon \in (0, 1]$), fix a target $\varepsilon' > 0$. If the no-new rule is run with temperatures bounded by*

$$\tau_t \;\le\; \frac{\varepsilon'}{\log(n_1^t/\varepsilon) + \log(n_2^t/\varepsilon)} \qquad (n_i^t = |\Pi_i^t|,\ \log(n_1^t/\varepsilon) + \log(n_2^t/\varepsilon) > 0),$$

then at the stopping round $\mathrm{NashConv}(\mu^t, \nu^t) \le \varepsilon'$, after at most $|S_1| + |S_2|$ rounds. The anchored QRE is therefore not confined to a fixed-approximation guarantee as a DO meta-solver: annealing the temperature toward zero recovers any prescribed equilibrium precision.

Proof. By (9) the no-new terminal exploitability is at most $\tau_t(\log(n_1^t/\varepsilon) + \log(n_2^t/\varepsilon)) \le \varepsilon'$, and the stopping-time bound is Proposition 5.4's. When the denominator vanishes ($\varepsilon = 1$ and $n_1^t = n_2^t = 1$) the stallgap bound is already $\mathrm{NashConv} = 0$. □ □

Proposition B.2 (The target rule)**.** *Run Algorithm 1 with $\varepsilon \in (0, 1]$, an exact best-response oracle, and the* target *rule with threshold $\theta > 0$: stop at the first round t with $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < \theta$, the gaps measured in the true game, and otherwise append every best response not already in its population. (i) Whenever the rule stops, $\mathrm{NashConv}(\mu^t, \nu^t) < 2\theta$, for exact and for estimated meta-games alike. (ii) With exact meta-games, at any round at which both exact best responses already lie in the populations, $\mathrm{gap}_i(\mu^t, \nu^t) \le \tau_t \log(n_i^t/\varepsilon)$ for each i. If the schedule satisfies $\tau_t \log(n_i^t/\varepsilon) < \theta$ for both players at every round, then every non-stopping round appends a new strategy, the rule stops after at most $|S_1| + |S_2|$ rounds, and it stops no later than the no-new rule run from the same start with the same schedule and the same best-response selections along the common history. (iii) With estimated meta-games, under the hypotheses of Corollary 6.2 and on its probability-$(1 - \delta)$ event, the conclusions of (ii) hold with the condition $\tau_t \log(n_i^t/\varepsilon) + 2\kappa_n < \theta$. When the schedule condition fails the rule may pass rounds at which both best responses are present yet a gap is still at least θ. At such a round nothing is appended and only the anchor moves, so termination is not covered by the argument below.*

Proof. (i) At a stopping round $\mathrm{NashConv} = \mathrm{gap}_1 + \mathrm{gap}_2 < 2\theta$ by the stopping test, which is evaluated in the true game whatever matrix the meta-solver saw. (ii) The per-player bound is the display in the proof of Proposition 5.4; it uses only that the best response lies in Π_i^t and the anchor floor $\rho_{i,s}^t \ge \varepsilon/n_i^t$. Under the schedule condition a round at which both best responses are present has $\max_i \mathrm{gap}_i < \theta$ and is a stopping round; every other round appends at least one new strategy, which happens at most $|S_1| + |S_2|$ times. When they do not stop, both rules append every new best response, and with the same selections along the common history the two trajectories coincide up to the earlier stopping time; the no-new rule stops at the first round with both best responses present, which is also a stopping round of the target rule. (iii) Replace the per-player bound by $\tau_t \log(n_i^t/\varepsilon) + 2\kappa_n$, established in part (ii) of the proof of Corollary 6.2 using only that the best response lies in Π_i^t; the probability event there covers $R = |S_1| + |S_2|$ rounds, which the target rule does not exceed under the condition. □ □

Remark B.3 (Stopping rules). Algorithm 1 states the target rule used by the experiments; Proposition 5.4 and Corollary 6.2 state their guarantees for the parameter-free no-new rule, which stops as soon as no new strategy is appended. The two rules differ in two regimes: when a new best response improves by less than the target, the target rule stops while the no-new rule appends it; or when all best responses are already in the populations yet one still improves by at least the target, the target rule continues while the no-new rule stops. Proposition B.2 rules the second regime out once the schedule condition holds. Whenever $\max_i \mathrm{gap}_i \ge 2/k$, at least one new strategy is appended under both rules; on G_k under the temperature cap this holds every round until the $2/k$ target, so the $4k$ bound of Corollary 5.5 holds under either.

Proof of Corollary 5.5. Corollary 5.2 is immediate from (14)–(15) as stated in the main text. For Corollary 5.5, verify the hypotheses of Theorem 4.3: with ε-mixing, every current-population strategy carries anchor mass $\rho_{i,s} \geq \varepsilon/\max(n_1, n_2) > 0$, so the anchors are interior (Definition A.1); moreover $\log(1/\rho^t_{\min}) \leq \log(\max(n_1, n_2)/\varepsilon) \leq \log(e\, n_1 n_2/\varepsilon)$ and condition (7) is implied by the stated temperature bound. □

Remark B.4 (ε is an exploration knob, not a numerical patch). Monotonicity in ε is proven for the mass floor (Lemma 5.3); it does *not* hold for final exploitability. The small-ε arm decreased at the three reported values (final exploitability $1.64/1.11/0.67$ as ε rises from 10^{-3} to 10^{-1}), but the full sweep is minimized at an interior $\varepsilon \approx 0.5$ and *worsens* again toward $\varepsilon = 1$ (Table 3 in Appendix E). So ε trades anchoring strength against exploration, with an interior optimum: the trade is neither monotone nor a binary switch.

C Proofs for Section 6

Corollary C.1 (Finite-sample meta-solutions)**.** *Let every entry of $\hat{A}$ be the average of n i.i.d. simulations with range $[-B, B]$ and mean A, and let both QREs use the same fixed temperature $\tau > 0$ and the same interior anchors. With probability at least $1 - \delta$,* (11) *holds. Moreover the exploitability of the computed profile degrades additively: for $\hat{z} = z^\star(\hat{A})$, $\mathrm{NashConv}_A(\hat{z}) \leq \mathrm{NashConv}_A(z^\star(A)) + 3B \cdot \frac{2\kappa_n}{\tau}$.*

Proof of Theorem 6.1. For interior anchors, the solutions $z^\star(A) = (\mu^\star, \nu^\star)$ and $\hat{z} = z^\star(\hat{A})$ are interior by (3), so the KL gradients below are finite. Define on $\Omega := \Delta(S_1) \times \Delta(S_2)$ the operator

$$F_A(\mu, \nu) := \big(-A\nu + \tau \nabla_\mu \mathrm{KL}(\mu\|\rho_1),\ \ A^\top \mu + \tau \nabla_\nu \mathrm{KL}(\nu\|\rho_2)\big).$$

Saddle-point optimality of $z^\star(A)$ for (2) is equivalent to the variational inequality $\langle F_A(z^\star), z - z^\star\rangle \geq 0$ for all $z \in \Omega$.

Strong monotonicity. For $z = (\mu, \nu)$, $z' = (\mu', \nu')$ in the interior,

$$\langle F_A(z) - F_A(z'),\, z - z'\rangle = \underbrace{\delta^\top A\eta - \delta^\top A\eta}_{=0} + \tau \sum_{i=1,2} \big(\mathrm{KL}(w_i\|w_i') + \mathrm{KL}(w_i'\|w_i)\big),$$

with $\delta := \mu - \mu'$ and $\eta := \nu - \nu'$; the two cross terms cancel. For $(w_1, w_2) = (\mu, \nu)$ one has $\langle \nabla \mathrm{KL}(w\|\rho) - \nabla \mathrm{KL}(w'\|\rho),\, w - w'\rangle = \sum_s (\log w_s - \log w_s')(w_s - w_s') = \mathrm{KL}(w\|w') + \mathrm{KL}(w'\|w)$, and by Pinsker's inequality in both directions $\mathrm{KL}(w\|w') + \mathrm{KL}(w'\|w) \geq \|w - w'\|_1^2$, so

$$\langle F_A(z) - F_A(z'),\, z - z'\rangle \ \geq\ \tau\big(\|\mu - \mu'\|_1^2 + \|\nu - \nu'\|_1^2\big) \ \geq\ \tfrac{\tau}{2}\, s^2, \qquad s := \|\mu - \mu'\|_1 + \|\nu - \nu'\|_1.$$

Perturbation. Apply the two variational inequalities at each other's solutions, $\langle F_A(z^\star), \hat{z} - z^\star\rangle \geq 0$ and $\langle F_{\hat{A}}(\hat{z}), z^\star - \hat{z}\rangle \geq 0$, and add:

$$\langle F_A(\hat{z}) - F_A(z^\star),\, \hat{z} - z^\star\rangle \ \leq\ \langle F_A(\hat{z}) - F_{\hat{A}}(\hat{z}),\, \hat{z} - z^\star\rangle.$$

The right-hand side equals $\langle (-(A - \hat{A})\hat{\nu},\, (A - \hat{A})^\top \hat{\mu}),\, \hat{z} - z^\star\rangle \leq \|(A - \hat{A})\hat{\nu}\|_\infty \|\hat{\mu} - \mu^\star\|_1 + \|(A - \hat{A})^\top \hat{\mu}\|_\infty \|\hat{\nu} - \nu^\star\|_1 \leq \|A - \hat{A}\|_\infty\, s$, since $\hat{\mu}, \hat{\nu}$ are probability vectors. Combining with strong monotonicity, $\frac{\tau}{2} s^2 \leq \|A - \hat{A}\|_\infty s$; if $s > 0$, dividing by s yields $s \leq \frac{2}{\tau}\|A - \hat{A}\|_\infty$ (the case $s = 0$ is trivial). □

Remark C.2 (Discontinuity of the Nash correspondence). For contrast, consider the 1×2 meta-game $A = [0, \delta]$: player 2's exact Nash support jumps from column 2 to column 1 as δ crosses 0, so arbitrarily close games can have Nash equilibria at ℓ_1-distance 2: the Nash solution correspondence has no continuous single-valued selection. Under meta-game noise of scale σ: if two Nash selections separated by $\Omega(1)$ occur on opposite sides of a restricted-game tie, and the noise law crosses that tie with probability bounded below, a Nash MSS resamples an $\Omega(1)$-different meta-solution with nonvanishing probability. For a common fixed anchor and $\|\hat{A} - A\|_\infty \leq \sigma$, the anchored QRE differs from the noiseless QRE by at most $2\sigma/\tau$. The $1/\tau$ rate of Theorem 6.1 is order-optimal: in this 1×2 example with symmetric anchors and temperature τ, $\nu_2(\delta) = 1/(1 + e^{\delta/\tau})$, so the profile slides at ℓ_1-speed $2\nu_1\nu_2/\tau \leq 1/(2\tau)$, attained at $\delta = 0$; no Lipschitz bound with an $o(1/\tau)$ rate can hold, whereas the Nash support here jumps by the full diameter 2.

Proof of Corollary C.1. Each entry of $\hat{A}$ is an average of n i.i.d. variables in $[-B, B]$ with mean A_{xy}; Hoeffding's inequality gives $\Pr[|\hat{A}_{xy} - A_{xy}| \geq t] \leq 2e^{-nt^2/(2B^2)}$, and a union bound over the $n_1 n_2$ entries yields $\|\hat{A} - A\|_\infty \leq \kappa_n$ with probability $\geq 1 - \delta$. The ℓ_1 bound is then Theorem 6.1. For the exploitability transfer, for any profiles $z = (\mu, \nu)$, $\hat{z} = (\hat{\mu}, \hat{\nu})$,

$$\big|u(\hat{\mu}, \hat{\nu}) - u(\mu, \nu)\big| \leq B\big(\|\hat{\mu} - \mu\|_1 + \|\hat{\nu} - \nu\|_1\big), \qquad \Big|\max_x u(x, \hat{\nu}) - \max_x u(x, \nu)\Big| \leq B\|\hat{\nu} - \nu\|_1,$$

so $\mathrm{gap}_1(\hat{z}) \leq \mathrm{gap}_1(z) + B(\|\Delta\mu\|_1 + 2\|\Delta\nu\|_1)$ and symmetrically for gap_2; summing, $\mathrm{NashConv}(\hat{z}) \leq \mathrm{NashConv}(z) + 3Bs$. Taking $z = z^\star$ (so $\mathrm{NashConv}(z) = \mathrm{NashConv}(z^\star)$) and bounding $s \leq 2\kappa_n/\tau$ by Theorem 6.1 gives $\mathrm{NashConv}(\hat{z}) \leq \mathrm{NashConv}(z^\star) + 3B \cdot \frac{2\kappa_n}{\tau}$. □

Proof of Corollary 6.2. *The probability event.* Let $\mathcal{H}_{t-1}$ denote everything known when the round-t populations, anchors, and temperature are set but before the round-t simulations are drawn. By the corollary's independence hypothesis the round-t simulations are fresh conditional on $\mathcal{H}_{t-1}$, so applying Corollary C.1 conditionally with confidence δ/R and the (then fixed) round-t sizes gives $\Pr[\|\hat{A}^t - A^t\|_\infty \leq \kappa_n \mid \mathcal{H}_{t-1}] \geq 1 - \delta/R$. The stopping-time bound of Proposition 5.4 caps the run at R rounds. Writing $\mathcal{E}_t := \{\|\hat{A}^t - A^t\|_\infty \leq \kappa_n\}$ (taken to hold trivially at rounds the run never reaches) and conditioning on $\mathcal{H}_{R-1}$, then on $\mathcal{H}_{R-2}$, and so on, $\Pr[\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_R] = \mathbb{E}\big[\mathbb{I}[\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_{R-1}] \Pr[\mathcal{E}_R \mid \mathcal{H}_{R-1}]\big] \geq (1 - \delta/R) \Pr[\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_{R-1}] \geq \cdots \geq (1 - \delta/R)^R \geq 1 - \delta$, which is the probability-$(1 - \delta)$ event of the corollary (with κ_n evaluated at δ/R and the round-t sizes). The two parts below hold on that event.

(i) The appended x^t maximizes $u(\cdot, \nu^t)$ (a best response in the true game to the current meta-solution $\nu^t = \nu^\star(\hat{A}^t)$). For the true target $\nu^\star = \nu^\star(A^t)$, using $\|\nu^t - \nu^\star\|_1 \leq 2\kappa_n/\tau$ (Theorem 6.1) twice,

$$u(x^t, \nu^\star) \geq u(x^t, \nu^t) - B\|\nu^t - \nu^\star\|_1 \geq \max_x u(x, \nu^t) - B\tfrac{2\kappa_n}{\tau} \geq \max_x u(x, \nu^\star) - \tfrac{4B\kappa_n}{\tau},$$

so x^t is a $4B\kappa_n/\tau_t$-approximate best response to the true target $\nu^\star$; symmetrically, y^t is a $4B\kappa_n/\tau_t$-approximate best response to $\mu^\star = \mu^\star(A^t)$.

(ii) At a stopping round t the no-new rule has the exact best response against ν^t lie in Π_1^t (and symmetrically), so $\max_{x \in S_1} u(x, \nu^t) = \max_{x \in \Pi_1^t} u(x, \nu^t) = \max_{x \in \Pi_1^t} e_x^\top A^t \nu^t$: the full best-response value against ν^t under the true matrix A^t is attained within Π_1^t. Fix the profile (μ^t, ν^t). By QRE optimality of μ^t against the matrix the solver actually solved, $\hat{A}^t$, comparing with the vertex at any restricted maximizer $x' \in \Pi_1^t$ of $x \mapsto e_x^\top \hat{A}^t \nu^t$ gives $\max_{x \in \Pi_1^t} e_x^\top \hat{A}^t \nu^t - \mu^{t\top} \hat{A}^t \nu^t \leq \tau_t \log(n_1/\varepsilon)$ (the anchor floor $\rho_{1,s}^t \geq \varepsilon/n_1$). Two fixed-profile matrix-change bounds, $|e_x^\top (A^t - \hat{A}^t)\nu^t| \leq \kappa_n$ for every x and $|\mu^{t\top}(A^t - \hat{A}^t)\nu^t| \leq \kappa_n$ (κ_n of (11)), then give

$$\mathrm{gap}_1^{A^t}(\mu^t, \nu^t) = \max_{x \in S_1} u(x, \nu^t) - \mu^{t\top} A^t \nu^t = \max_{x \in \Pi_1^t} e_x^\top A^t \nu^t - \mu^{t\top} A^t \nu^t \;\leq\; \tau_t \log \tfrac{n_1}{\varepsilon} + 2\kappa_n,$$

where the first equality uses the no-new coincidence (the full maximum is attained within Π_1^t). Adding the symmetric player-2 bound gives (12). □

Corollary C.3 (Estimation noise does not reintroduce the worst case)**.** *Run DO on G_k from $\Pi_1^0 = \{1^k\}$, $\Pi_2^0 = \{0^k\}$ exactly as in Theorem 4.3: interior anchors ρ^t, temperatures satisfying* (7)*, an exact best-response oracle under arbitrary adversarial tie-breaking, terminating only when $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ strictly. Suppose only that the restricted game the meta-solver solves at round t is the* estimate *$\hat{A}^t$: every entry is the average of n fresh i.i.d. simulations with range $[-B, B]$ and mean the true entry (on G_k every payoff satisfies $|u| \leq 1$, so $B = 1$). Let $R = 4k$. If the sample size obeys*

$$n \;\geq\; 32\, B^2 k^2 \,\log \frac{2R^3}{\delta}, \tag{18}$$

then with probability at least $1 - \delta$ the run reaches $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ within at most $R = 4k$ rounds.

Proof. The probability event. Write $\kappa_R := B\sqrt{\frac{2}{n} \log \frac{2R^3}{\delta}}$, so that (18) implies $2\kappa_R \leq \frac{1}{2k}$. Each population starts as a singleton and grows by at most one strategy per round, so at each of the first R

rounds ($t = 0, \dots, R-1$) $|\Pi_i^t| \le t+1 \le R$ and the round-t estimate has at most R^2 entries; Hoeffding on every entry ((11)) with a union bound over the entries gives, conditional on the round-t history (the draws are fresh, exactly as in the proof of Corollary 6.2), $\Pr[\|\hat{A}^t - A^t\|_\infty \le \kappa_R] \ge 1 - \delta/R$ for every one of the at most R rounds. Iterated conditioning over the rounds, as in that proof, bounds the joint event by $(1-\delta/R)^R \ge 1-\delta$. The rest of the proof runs on the event $\{\forall\, t < R,\ \|\hat{A}^t - A^t\|_\infty \le \kappa_R\}$, on which no further probability is spent.

Step 1 under noise. Say $x \in \Pi_1^t$ is a 1-achiever at round t if $u(x, \nu^t) = \max_{x' \in \{0,1\}^k} u(x', \nu^t)$, and define 2-achievers symmetrically, exactly as in the proof of Theorem 4.3 (best responses in the true game to the returned profile). If a 1-achiever $x^\star$ lies in Π_1^t, then the suboptimality of μ^t against $x^\star$ in the matrix $\hat{A}^t$ that the meta-solver actually solved is at most $\tau_t \log(1/\rho_{\min}^t)$ by the KL bound of the exact proof, and decomposing the true payoff around the estimate, with $E^t := \hat{A}^t - A^t$, gives

$$\begin{aligned} \mathrm{gap}_1(\mu^t, \nu^t) &= u(x^\star, \nu^t) - \mu^{t\top} A^t \nu^t \\ &= \big(x^{\star\top} \hat{A}^t \nu^t - \mu^{t\top} \hat{A}^t \nu^t\big) + \mu^{t\top} E^t \nu^t - x^{\star\top} E^t \nu^t \\ &\le \tau_t \log \tfrac{1}{\rho_{\min}^t} + \big|\mu^{t\top} E^t \nu^t\big| + \big|x^{\star\top} E^t \nu^t\big| \\ &\le \tau_t \log \tfrac{1}{\rho_{\min}^t} + 2\kappa_R \ \le\ \tfrac{1}{k} + \tfrac{1}{2k} = \tfrac{3}{2k} \ <\ \tfrac{2}{k}, \end{aligned} \tag{19}$$

where the last line uses (7), the event bound $\|E^t\|_\infty \le \kappa_R$, and $\|\mu^t\|_1 = \|x^\star\|_1 = 1$. Symmetrically, the presence of a 2-achiever caps $\mathrm{gap}_2 \le 3/(2k) < 2/k$. Hence at a *blocked* round ($\max_i \mathrm{gap}_i \ge 2/k$) at least one of the two achiever sets is still empty, exactly as in the exact proof.

Steps 2–4 (charging) are unchanged. The remaining argument bounds the number of 1-blocked and 2-blocked rounds using only: interiority of the meta-solver (the anchored QRE of $\hat{A}^t$ is interior for every matrix when the anchors are interior and $\tau_t > 0$, by (3)), exactness of the best responses in the true game, and the forcing structure of dirty and attacked bits (Lemma A.2), none of which refers to the matrix the meta-solver solved. So at most $(2k-1) + 2k = 4k - 1 = R - 1$ blocked rounds can precede termination, and on the event above the run reaches $\max_i \mathrm{gap}_i(\mu^t, \nu^t) < 2/k$ within at most $R = 4k$ rounds. □ □

Corollary C.4 (Annealing reaches any precision under meta-game noise)**.** *Run Algorithm 1 under the hypotheses of Corollary 6.2 ($\varepsilon \in (0, 1]$, n fresh simulations per entry per round, an exact best-response oracle, full-game payoffs in $[-B, B]$, the no-new stopping rule, hence at most $R = |S_1| + |S_2|$ rounds). Fix a target $\varepsilon' > 0$. If the temperatures satisfy*

$$\tau_t \ \le\ \frac{\varepsilon'}{2\big(\log(n_1^t/\varepsilon) + \log(n_2^t/\varepsilon)\big)} \qquad (n_i^t = |\Pi_i^t|,\ \textit{whenever the denominator is positive}) \tag{20}$$

and the sample size satisfies

$$n \ \ge\ \frac{128\, B^2}{\varepsilon'^2} \log \frac{2R^3}{\delta}, \tag{21}$$

then with probability at least $1-\delta$ the stopping round t has $\mathrm{NashConv}(\mu^t, \nu^t) \le \varepsilon'$. The perturbation bias and the estimation term of (12) *are thus driven below $\varepsilon'/2$ each by one schedule and one sample size. The two terms add, so as $n \to \infty$ the requirement* (21) *disappears and the schedule can be relaxed to that of Corollary B.1.*

Proof. On the probability-$(1-\delta)$ event of Corollary 6.2 (confidence δ/R per round), (12) holds at the stopping round: $\mathrm{NashConv}(\mu^t, \nu^t) \le \tau_t L_t + 4\kappa_n$ with $L_t = \log(n_1^t/\varepsilon) + \log(n_2^t/\varepsilon)$ and, since $n_1^t n_2^t \le R^2$, $\kappa_n \le B\sqrt{\frac{2}{n} \log \frac{2R^3}{\delta}}$. The schedule (20) gives $\tau_t L_t \le \varepsilon'/2$ (when $L_t = 0$ the bias term is already 0), and (21) gives $4\kappa_n \le 4B\sqrt{\frac{2}{n} \log \frac{2R^3}{\delta}} \le 4B \cdot \frac{\varepsilon'}{8B} = \varepsilon'/2$. The conditions of Corollary 6.2 carry over: the guarantee is conditional on exact best responses and exact QRE computation. □ □

D SUPPLEMENTAL EXPERIMENTS

Deadlock three-branch verification. The deadlock study of Section 7 exercises all three branches of Theorem 5.1 on the 60×60 Gaussian matrix with the anchor set to the previous round's meta-solution, $\varepsilon = 0$, $\tau = 0.5$, 40 rounds, and two initial configurations. The *singleton* configuration (10

seeds) checks the freezing branch: the meta-solution must be identical at every round and exploitability constant. The *two-strategy* configuration (10 seeds) checks the in-support bound: $\mathrm{NashConv}_t$ must never fall below the minimum NashConv achievable on the initial supports (computed exactly by two linear programs whose variables range over the initial supports while the deviation constraints range over all 60 full-game strategies; a restricted 2×2 Nash solve would give the trivial bound 0), and the maximum weight on newly added strategies must be below 10^{-6} (support trapping). All criteria were fixed before the runs and all passed on every seed (the aggregation script checks the stricter 10^{-290} for the new-strategy weight; Table 2).

Table 2: Three-branch verification of Theorem 5.1 at $\varepsilon = 0$ (10 seeds per branch).

branch	initial supports	criterion	outcome
freezing	singleton	meta-solution identical, NashConv constant	10/10
in-support bound	two strategies	$\mathrm{NashConv}_t \geq \min$ NC on initial supports	10/10
support trapping	two strategies	max new-strategy weight $< 10^{-6}$	10/10

Table 3 reports the ε-repair sweep of Section 7.

Table 3: Anchoring deadlock and ε-repair on the 60×60 matrix (10 seeds). "Hard deadlock": stalled exploitability and maximum new-strategy weight below 10^{-6} after round 5. Final exploitability is minimized at an interior $\varepsilon = 0.5$; the $\varepsilon \geq 0.3$ rows replay neither the deadlock criterion nor the new-strategy weight (–).

ε	hard deadlocks	median max new-strategy weight	mean final expl.	std
0 (raw)	10/10	0	4.46	–
10^{-3}	0/10	5.4×10^{-2}	1.637	0.333
10^{-2}	0/10	1.4×10^{-1}	1.106	0.160
10^{-1}	0/10	2.4×10^{-1}	0.672	0.133
0.3	–	–	0.298	0.055
0.5	–	–	0.253	0.029
1.0	–	–	0.406	0.066

In the two-strategy runs the in-support NashConv bound holds with margin 0.04–1.44 across all rounds and seeds. The trapped meta-solution drifts within the initial supports, with per-seed NashConv range 0.13–1.47 in the multiplicative-weights motion identified in the proof of Theorem 5.1; the maximum weight ever placed on a new strategy is below 10^{-296} on every seed.

Anchor-value ablation under meta-game noise. Section 7 reports the exact-meta-game half of the anchor-value sweep (Table 3). The noisy half re-samples every meta-game entry from scratch each round (no reuse across rounds) with $n \in \{50, 200\}$ simulations, 10 seeds on the 60×60 matrix and Kuhn poker and 5 seeds on Leduc. Table 4 reports the mean final exploitability over the ε grid. The interior optimum $\varepsilon = 0.5$ is best at $n = 50$ on the matrix and Kuhn (paired difference against $\varepsilon = 1.0$: -0.11 ± 0.13 and -0.02 ± 0.05), while $\varepsilon = 1.0$ is best at $n = 200$ ($+0.14 \pm 0.15$ in its favor on the matrix, within 0.003 on Kuhn); on Leduc the best ε varies with n (0.3 at $n = 50$, 0.01 at $n = 200$) within a run-to-run noise band larger than the differences. None of the matrix and Kuhn intervals excludes zero, so the noisy runs are inconclusive for the interior optimum. The anchor's advantage concentrates on the exact meta-game (-0.154 ± 0.037, Table 3).

Terminal finite-sample bound. On a 30×30 matrix whose entries are uniform on $[-0.9, 0.9]$, with each simulation a $\{\pm 1\}$ Bernoulli draw of the entry (range $B = 1$; Appendix F), we set $\tau = 0.05$, $\varepsilon = 0.1$, confidence $\delta = 0.1$, 10 seeds, and n fresh simulations per entry each round. We run Algorithm 1 under two stopping rules, the improvement-threshold rule used in the main text and the no-new-strategy rule of Proposition 5.4, and compare the terminal NashConv against the bound (12) (Table 5). The joint event of Corollary 6.2 (per-round concentration plus its union bound over rounds) holds on all 80 runs, and no run's terminal NashConv violates its bound. The bound is loose by a factor of 6–33, dominated by the union bound in its derivation.

Table 4: Mean final exploitability under fresh noise, by (domain, n, ε); lower is better. †: Leduc runs use 5 seeds.

domain	n	ε=10^{-3}	10^{-2}	10^{-1}	0.3	0.5	1.0
matrix	50	0.849	0.878	0.802	0.819	0.749	0.863
matrix	200	0.829	0.805	0.801	0.767	0.830	0.688
Kuhn	50	0.205	0.188	0.155	0.167	0.125	0.148
Kuhn	200	0.122	0.079	0.091	0.076	0.079	0.076
Leduc†	50	2.816	2.721	2.579	2.288	2.723	2.773
Leduc†	200	2.700	2.580	2.786	2.720	2.967	2.633

Corollary 6.2 is stated for the no-new-strategy stopping rule; for the improvement rule (threshold $\theta = 0.05$) the same bound is applied as an empirical upper envelope, and it is never violated either, while the guarantee that does apply, Proposition B.2(i), holds on all 40 runs (terminal $\mathrm{NashConv} \leq 0.094 < 2\theta$). The termination condition of Proposition B.2(ii) fails here ($\tau \log(n_i/\varepsilon) \geq 0.149 > \theta$ already at $n_i = 2$), and the uncovered regime does occur: 33 of the 40 runs pass a total of 60 rounds at which both best responses are present yet a gap exceeds θ; the anchor iteration then closes the gap without appending, and every run stops. The improvement rule terminates with terminal $\mathrm{NashConv}$ 44–59% below the no-new-strategy rule at every budget: stopping on a target improvement stays inside the envelope and is empirically *better* than the parameter-free rule of Proposition 5.4, the conservative arm.

Table 5: Terminal bound exercise (10 seeds per cell). "Event": the joint event of Corollary 6.2 holds; "violations": terminal $\mathrm{NashConv}$ above its bound.

n	rule	stopped	event	violations	$\overline{\mathrm{NashConv}}_{\mathrm{term}}$	bound/NashConv	stop round
100	improvement	10	10	0	0.079	33.4	24.9
100	no-new	10	10	0	0.141	21.4	20.8
400	improvement	10	10	0	0.063	25.5	22.5
400	no-new	10	10	0	0.134	13.4	19.7
1600	improvement	10	10	0	0.066	16.8	21.5
1600	no-new	10	10	0	0.125	12.1	19.7
6400	improvement	10	10	0	0.057	14.7	21.6
6400	no-new	10	10	0	0.140	6.5	18.4

Lipschitz micro-study. For three temperatures $\tau \in \{0.05, 0.2, 1.0\}$, an 11-point geometric scale of payoff perturbations $\sigma \in [10^{-4}, 0.3]$, and 10 noise seeds, we compare the ℓ_1-movement of the anchored-QRE meta-solution against the bound $2\sigma/\tau$ of Theorem 6.1. All 330 configurations satisfy the bound (violations 0); the tightest uses $\approx 10\%$ of the bound (median 6%). The exact-Nash solution correspondence, in contrast, jumps by the full ℓ_1-diameter 2 when a restricted-equilibrium tie is crossed (Figure 3).

Approximate best-response oracle. Section 7 reports the headline outcome; here we give the protocol and full numbers. On Leduc poker we replace the exact best-response oracle of Algorithm 1 by a deep Q-network (Mnih et al., 2015) trained against the current opponent mixture for 60k environment steps per round (single hidden layer, fixed random seeds; training is deterministic under each seed). Three meta-solvers (anchored QRE, PRD, RRD) $\times$ three seeds, 8 rounds, 100 simulations per meta-game entry. Table 6 reports the final $\mathrm{NashConv}$ (evaluated on the full population after the last addition) and the mean oracle gap, the shortfall of the trained policy against the exact best response to the same opponent mixture. Every added policy is near-deterministic (no exact duplicate is ever produced). The oracle gap grows with the complexity of the opponent mixture (round 1: $\approx$ 0.1–0.4; later rounds: up to $\approx$ 1.9) and never falls back, so the approximate oracle does not converge to exact behavior within the run.

The final-$\mathrm{NashConv}$ differences between meta-solvers (0.17–0.71) are comparable to the oracle gap (0.53–0.75), so the ranking (RRD below anchored QRE below PRD) is not statistically resolvable at three seeds (both paired bootstrap intervals involving the anchored QRE cover zero; only RRD

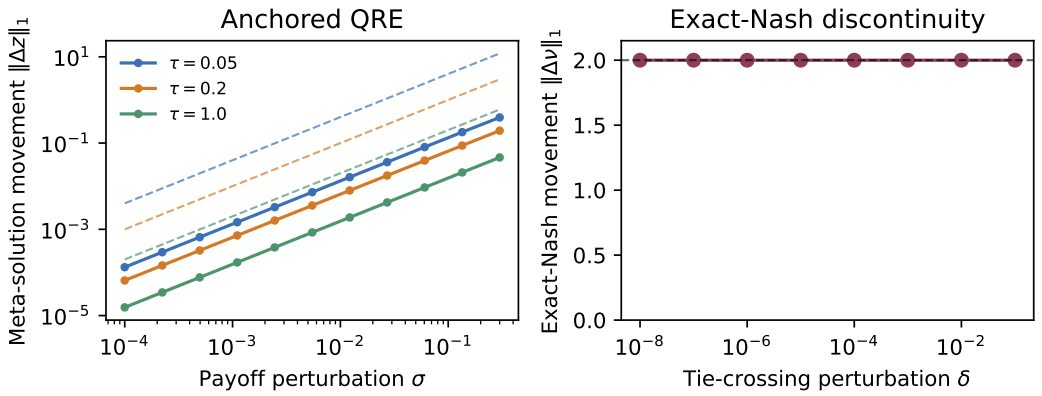


Figure 3: (a) Anchored-QRE meta-solution ℓ_1-movement against payoff perturbation σ for $\tau \in \{0.05, 0.2, 1.0\}$ (solid: measured; dashed: the bound $2\sigma/\tau$ of (10)): the movement grows linearly in σ and shrinks with $1/\tau$, staying below the bound throughout. (b) The exact-Nash solution correspondence is discontinuous: crossing a tie by δ flips the support and the solution jumps by ℓ_1-diameter 2 however small δ is.

Table 6: Deep Q-network (DQN) oracle PSRO on Leduc (3 seeds). Final NashConv is evaluated after the last strategy is added.

meta-solver	final NashConv (mean $\pm$ SD)	mean oracle gap
anchored QRE	2.441 ± 0.406	0.646
PRD	2.984 ± 0.680	0.599
RRD	2.274 ± 0.135	0.632

against PRD, -0.71 $[-1.61, -0.25]$, excludes it). With an approximate oracle of this accuracy, meta-solver differences of the size our theory predicts are not reliably measurable at three seeds. The qualitative deadlock behaviour of the exact case does survive: in every run the population grows to 9 strategies per player with no exact duplicate, i.e. the ε-mixed anchor keeps adding genuinely new (near-deterministic) strategies throughout.

η-approximate best responses on G_k. Theorem A.7 bounds the escape under η-approximate best responses by the QRE's minimum mass, (16). Under the theorem's novelty-constrained oracle the run stays below the bound as in the exact case, the admissible budget shrinking with k (the $e^{-2/\tau}$ factor of (13), Table 9). A far larger budget collapses the mechanism on small k and stretches the run on larger ones. Without the append rule the same oracle stalls immediately, freezing at $2/k$ for any $\eta \geq 1$ (Table 10), so the append rule is necessary against large-error oracles (Remark A.8). At the probe's capped temperatures (16) falls below 10^{-21} from round two, while delays appear only for $\eta \gtrsim 10^{-1}$ (Remark A.10).

E Supporting figures and tables

This appendix collects the tables referenced from the main text, together with the protocol details they instantiate.

Escape probes: mixed-Nash-uniform, α-DO, and η-approximate BRs. Table 8 reports the rounds to gap $< 2/k$ behind Section 7. Probe P1 (exact Nash mixed with an ε-uniform term, adversarial ties) escapes on every k in 3 rounds, so full support alone suffices; its $\varepsilon = 10^{-2}$ satisfies the condition $\varepsilon \leq 1/(4k)$ of Corollary A.9 at every tested k, so the probe is an instance of that corollary. Probe P2 (an α-double-oracle selection (Zhang & Sandholm, 2024), a best response added only when its value improvement is at least α) escapes for $\alpha \in \{0.01, 0.1\}$ at every k and at $k = 4$ for $\alpha = 0.5$, where $2/k = \alpha$ exactly. It stalls immediately for $\alpha = 0.5$ on every $k \geq 5$, where $\alpha > 2/k$.

Table 7: On G_k: groups B–E report rounds to gap $< 2/k$ (median $\pm$ half-range over seeds; deterministic groups 3 seeds, stochastic 10); group A reports the rounds the adversarial schedule of Zhang & Sandholm (2024) pins the gap at exactly $2/k$, namely $2^{k-1} - 1$ (one further round drops it below $2/k$).

group	meta-solver / tie-break	k=8	k=11	k=14
A	adversarial NE / adversarial	127	1023	8191
B	QRE-U / adversarial	4 ± 0	4 ± 0	5 ± 0
C	adversarial NE / uniquified	4 ± 0	4 ± 0.5	4 ± 0.5
D	QRE-U / random	3 ± 1	4 ± 1	4 ± 1.5
E	LP-NE / random	3.5 ± 1	4 ± 1	4 ± 3

Table 9 gives the η-approximate oracle sweep behind Appendix D: rounds until gap $< 2/k$ with adversarial tie-breaking toward the worst legal action, and the peak gap_2 in parentheses; the oracle is the novelty-constrained oracle assumed by Theorem A.7, returning a best response only when it is not already in the population. The admissible budget shrinks with k (delay appears at $\eta^\star \in (1, 10]$ for $k = 6$, at $\eta^\star \approx 10^{-1}$ for $k = 10$), and a budget far above the bound ($\eta = 10$) collapses player 2's revealing best responses on $k \leq 5$ (gap_2 near 0) while stretching the run to 10–18 rounds on larger k.

Table 10 drops the append rule: the same oracle may return a strategy already in the population. With that change, any $\eta \geq 1$ freezes the run at the initial gap $2/k$ on every k ("stall": none of the 3 seeds converged within 200 rounds), so the append rule is necessary against large-error oracles; Remark A.8 gives the matching pen-and-paper failure for any $\eta \geq 2/k$. For scale, the sufficient budget of (16) along the probe's own schedule ($\varepsilon = 10^{-2}$, $\tau_t = \min\{2/(t+1), [k \log(1/\rho^t_{\min})]^{-1}\}$) is $2e^{-2}/k \approx 0.03$–0.07 in the first round, where the anchor is the singleton point mass, and below 10^{-21} from the second round on, where $\rho^t_{\min} = \varepsilon/|\Pi^t_i|$ caps τ_t at about 0.02–0.05; the probe's delays at $\eta \gtrsim 10^{-1}$ therefore sit far above the theorem's guarantee.

Table 8: Rounds to gap $< 2/k$ on G_k (median; "stall": none of the 3 seeds converged). Probe P1 = exact-Nash mixed with an ε-uniform term, probe P2 = α-double-oracle. The stall at $k \geq 5$ for $\alpha = 0.5$ coincides with $\alpha > 2/k$; at $k = 4$, where $\alpha = 2/k$, the probe still escapes.

probe	meta-solver / selection	k=6	k=10	k=14
P1	NE + uniform, adversarial ties	3 (3/3)	3 (3/3)	3 (3/3)
P2, $\alpha = 0.01$	α-DO	3 (3/3)	3 (3/3)	3 (3/3)
P2, $\alpha = 0.1$	α-DO	3 (3/3)	3 (3/3)	3 (3/3)
P2, $\alpha = 0.5$	α-DO	stall (0/3)	stall (0/3)	stall (0/3)

Table 9: η-approximate best responses on G_k under the anchored QRE with adversarial tie-breaking toward the worst legal action and the novelty constraint of the double-oracle append rule. Entries: median rounds until gap $< 2/k$, and the peak gap_2 in parentheses. η above the bound (16) delays the escape; a budget far above it collapses player 2's revealing best responses on small k.

k	η=0	10^{-3}	10^{-2}	10^{-1}	1	10
4	3 (0.62)	3 (0.62)	3 (0.62)	3 (0.62)	3 (0.62)	2 (0.01)
5	3 (0.70)	3 (0.70)	3 (0.70)	3 (0.70)	3 (0.70)	2 (0.01)
6	3 (0.75)	3 (0.75)	3 (0.75)	3 (0.75)	3 (0.75)	18 (1.12)
7	3 (0.79)	3 (0.79)	3 (0.79)	3 (0.79)	3 (0.79)	10 (1.17)
8	3 (0.81)	3 (0.81)	3 (0.81)	3 (0.81)	3 (0.81)	18 (1.20)
9	3 (0.83)	3 (0.83)	3 (0.83)	3 (0.83)	3 (0.83)	11 (1.23)
10	3 (0.85)	3 (0.85)	3 (0.85)	4 (0.85)	4 (0.85)	18 (1.26)

Exact-to-noisy degradation and output variance. Table 11 supports Section 7: all eight meta-solvers are run on a near-exact meta-game and compared with the $n = 100$ runs of Table 1. The near-exact meta-game is the true payoff matrix on the random matrix and a 10,000-episode estimate

Table 10: The same η-approximate oracle without the double-oracle append rule, free to return a strategy already in the population. Any $\eta \geq 1$ freezes the run at the initial gap $2/k$ ("stall": none of the 3 seeds converged within 200 rounds); only exact best responses ($\eta = 0$) still escape. This isolates the append rule as the carrier of the escape.

k	η=0	η=1	η=10
4	3 (3/3)	stall (0/3)	stall (0/3)
6	3 (3/3)	stall (0/3)	stall (0/3)
8	3 (3/3)	stall (0/3)	stall (0/3)
10	4 (3/3)	stall (0/3)	stall (0/3)

per entry on the OpenSpiel domains (Lanctot et al., 2019), whose per-entry standard deviation is 0.009–0.045 across the domains, a factor $\sqrt{10}$ smaller than the $n = 1000$ noise level everywhere; "exact" in the tables below means this high-sample estimate. Table 13 puts the matrix exact and noisy runs on one axis: since an exact run consumes no simulations, its AUC is taken over the 25 rounds (mean exact exploitability per round), and the noisy runs are scored the same way on the same ten seeds.

Table 14 reports the fixed-population probe: on a population reached by the anchored QRE, the meta-game entries are re-estimated repeatedly at n episodes and the standard deviation of each solver's output exploitability over the re-estimates is listed. Exact Nash, maximum-entropy Nash and NE+uniform were all run; their output spreads are all within 0.002 of exact Nash's on every cell (NE+uniform mixes ε of the uniform into the Nash output and maximum-entropy Nash picks one equilibrium of the same LP), so only the first is shown; the uniform-anchored QRE (QRE-U) represents the QRE regularization at the paper's default temperature schedule. Together the two tables establish the noisy ranking: the near-exact meta-game preserves the mixed-domain wins, and the QRE regularization cuts the Nash solver's output spread on every cell.

Table 11: Difference between reported terminal means, Δ = final-expl(n=100) − final-expl(exact), mean over seeds (10 noisy and 10 exact seeds on the matrix, Kuhn, Goofspiel-4 and dark hex; 10 noisy and 5 exact seeds on Leduc, Goofspiel-5 and liar's dice, where the pools differ). Positive Δ: the noisy run ends worse. The two largest differences are on Leduc, NE+U at $+0.415$ and AQRE at $+0.401$; only the anchored QRE and PRD exceed 3.4 on more than one of the ten noisy Leduc seeds (3 each, against exact maxima of 3.21 and 3.12). Several entries are negative, so noise does not uniformly raise terminal exploitability; the near-exact AUC of every solver is in Table 12.

domain	NE	NE+U	MAXENT	PRD	RRD	α-Rank	QRE-U	AQRE
matrix (30×30)	$+0.133$	$+0.126$	$+0.133$	-0.031	$+0.055$	$+0.072$	$+0.081$	$+0.092$
Kuhn poker	$+0.132$	$+0.126$	$+0.125$	$+0.129$	$+0.038$	$+0.032$	$+0.113$	$+0.134$
Leduc poker	$+0.082$	$+0.415$	-0.452	$+0.196$	-0.229	$+0.096$	-0.231	$+0.401$
Goofspiel-4	-0.083	0.000	-0.113	$+0.068$	-0.002	$+0.014$	$+0.002$	-0.004
Goofspiel-5	$+0.036$	-0.132	$+0.148$	$+0.225$	$+0.013$	-0.197	-0.281	$+0.046$
dark hex (2×2)	0.000	0.000	0.000	0.000	$+0.006$	0.000	0.000	0.000
liar's dice (1d4)	$+0.055$	$+0.065$	$+0.030$	-0.113	-0.165	-0.027	$+0.037$	$+0.018$

Absolute results and fixed comparators. Table 1 compares the anchored QRE against the strongest baseline of each cell, chosen on the same data, so its intervals carry a selection bias in the baseline's favor; the liar's dice $n = 1000$ cell against α-Rank has raw paired interval $[+0.002, +0.038]$, a nominal loss excluding zero by 0.002. A five-seed pilot had already shown the same Leduc loss at $n = 10$ ($+0.16$); adding seeds narrowed the interval without moving the mean.

Table 15 gives the absolute AUC of all eight solvers in every cell, and Table 16 the paired differences against two comparators fixed in advance: RRD, the strongest replicator baseline, and the ε-mixed Nash NE+U, the rule of Corollary A.9. The bootstrap intervals are cellwise; for family-level evidence each cell also carries a paired t-test whose p-value is Holm-corrected over the 21 cells of its column.

Against RRD the anchored QRE is better in 15 of 21 cells (matrix, Goofspiel-4/5, Leduc and liar's dice at every n) and worse in 6 (Kuhn and dark hex at every n); 10 of the gains (Goofspiel-4/5 and

Table 12: Near-exact meta-game AUC (10,000 episodes per entry) by solver on the OpenSpiel domains (lower better; AUC is undefined on the matrix, whose exact exploitability is a point value). Best entry per domain in bold: the anchored QRE on Leduc and Goofspiel-4/5, the uniform-anchored QRE on Kuhn and dark hex, α-Rank on liar's dice, the same best-per-domain map as the $n = 100$ rows of Table 15.

domain	NE	NE+U	MAXENT	PRD	RRD	α-Rank	QRE-U	AQRE
Kuhn poker	0.530	0.529	0.530	0.600	0.500	0.562	**0.490**	0.567
Leduc poker	5.345	5.049	5.111	5.082	4.999	5.135	4.962	**4.875**
Goofspiel-4	1.257	1.416	1.117	1.401	1.159	1.373	1.132	**1.019**
Goofspiel-5	1.843	1.898	1.875	1.881	1.840	1.853	1.817	**1.712**
dark hex (2×2)	0.394	0.628	0.611	0.579	0.374	0.450	**0.323**	0.727
liar's dice (1d4)	1.405	1.510	1.466	1.428	1.420	**1.292**	1.367	1.302

Table 13: Matrix (30×30), exact versus noisy meta-game on one axis: mean exact exploitability over the 25 rounds (round-axis AUC), mean over the same ten seeds in every row. In parentheses: the paired difference to exact Nash and how many seeds the solver leads on. The anchored QRE is even with exact Nash on the exact meta-game and ahead of it under noise at every n; the uniform-anchored QRE has the lowest mean in every regime, so on this axis the matrix is a QRE-U domain, and the cumulative-simulation axis of Table 15 (on which the anchored QRE is best under noise) weights the two solvers' budgets differently.

solver	exact	$n = 10$	$n = 100$	$n = 1000$
NE	1.282	1.369	1.373	1.289
NE+U	1.279 (−0.003, 9)	1.365 (−0.004, 9)	1.371 (−0.002, 9)	1.286 (−0.003, 8)
MAXENT	1.282 (0.000, 5)	1.369 (0.000, 5)	1.373 (0.000, 7)	1.289 (0.000, 3)
PRD	1.495 (+0.213, 0)	1.541 (+0.172, 0)	1.419 (+0.046, 2)	1.497 (+0.208, 0)
RRD	1.252 (−0.030, 7)	1.286 (−0.084, 9)	1.253 (−0.120, 10)	1.249 (−0.040, 7)
α-Rank	1.348 (+0.066, 6)	1.396 (+0.027, 7)	1.403 (+0.030, 6)	1.286 (−0.003, 5)
QRE-U	**1.187** (−0.095, 8)	**1.277** (−0.092, 9)	**1.211** (−0.162, 9)	**1.178** (−0.111, 8)
AQRE	1.274 (−0.008, 5)	1.329 (−0.040, 7)	1.297 (−0.076, 8)	1.269 (−0.020, 6)

liar's dice at every n, Leduc at $n = 10$) and 5 of the losses (all but Kuhn at $n = 10$) survive Holm correction, the matrix gains ($p = 0.02$–0.04 cellwise) do not. Against NE+U it is better in 15, worse in 4 (dark hex at every n, Kuhn at $n = 1000$) and not significantly different in 2; all 15 gains and 2 of the 4 losses (Kuhn at $n = 1000$, dark hex at $n = 100$) survive Holm correction.

The tables also isolate the losses: on both Kuhn and dark hex the uniform-anchored QRE QRE-U, run with the same temperature schedule, has the lowest mean AUC of the eight solvers at every n, while the final exploitability (Table 18) of the anchored QRE is zero on dark hex, as for every solver except RRD and NE+U, and trails the best solver's by at most 0.035 on Kuhn; its paired terminal deficits against that best solver are $+0.021$ $[-0.064, +0.116]$ at $n = 10$, $+0.033$ $[+0.004, +0.064]$ at $n = 100$, and $+0.018$ $[-0.004, +0.043]$ at $n = 1000$, so only the $n = 100$ deficit excludes zero (95% paired bootstrap over the ten seeds), and on dark hex it is zero on every seed. The AUC deficit is therefore mainly one of speed, and the ablation identifies the last-iterate anchor as its source.

The mechanism is slow adoption, and Table 19 measures it: re-running the two QRE solvers with the meta-solution weights traced, the mass placed on the best responses appended in the previous round averages 0.08–0.14 on Kuhn and 0.02–0.03 on dark hex under the anchored QRE, against 0.44–0.53 for QRE-U with the same schedule, a 4–18× gap that holds on every seed and at every n; the median anchored mass on dark hex, 0.003, sits at the anchor floor ε/n_i to three decimals. The lower uptake of newly added strategies under the anchor (mass ε/n_i on a new entry, pulled toward the previous meta-solution with weight τ) is consistent with the AUC deficit: it is anchor propagation, the quantity Corollary 6.2 holds fixed per round, and a mild form of the deadlock of Theorem 5.1. The hot-schedule deficit of the ablation on Kuhn has the same source, a larger τ pulling harder toward the anchor.

Table 14: Fixed-population output variance (standard deviation of exploitability over repeated meta-game re-estimates). QRE-U is the QRE with a *fixed* uniform anchor and the default temperature schedule; the last-iterate anchor is meaningless across independent re-estimates of the same population, so this column isolates the QRE regularization itself. Every solver's spread falls with n at roughly $1/\sqrt{n}$; the QRE cuts the Nash solver's spread by about 2–5$\times$ on every cell and is the most stable solver on Kuhn, consistent with the smoothing intuition of Section 1.

domain	n	NE	RRD	α-Rank	QRE-U
matrix	10	0.102	0.083	0.023	0.049
matrix	100	0.035	0.030	0.010	0.014
matrix	1000	0.011	0.015	0.003	0.005
Kuhn	10	0.134	0.129	0.069	0.046
Kuhn	100	0.070	0.061	0.033	0.015

Table 15: Absolute normalized AUC of exact exploitability (mean over seeds; lower is better) for all eight meta-solvers in every cell of the controlled-noise study; the best entry per row is in bold. NE+U is exact Nash mixed with $\varepsilon = 10^{-2}$ uniform, MAXENT the maximum-entropy Nash, QRE-U the QRE with a fixed uniform anchor and the default schedule; 10 seeds per cell.

domain	n	NE	NE+U	MAXENT	PRD	RRD	α-Rank	QRE-U	AQRE
matrix (30×30)	10	2.631	2.633	2.631	2.769	2.489	2.733	2.548	**2.302**
	100	2.575	2.555	2.575	2.663	2.463	2.830	2.417	**2.274**
	1000	2.494	2.484	2.494	2.647	2.425	2.704	2.392	**2.246**
Kuhn poker	10	0.703	0.696	0.698	0.682	0.665	0.675	**0.663**	0.698
	100	0.592	0.592	0.592	0.617	0.533	0.607	**0.525**	0.618
	1000	0.542	0.541	0.542	0.587	0.501	0.589	**0.497**	0.586
Leduc poker	10	4.928	4.899	4.882	4.909	4.946	**4.490**	4.791	4.649
	100	5.305	5.125	5.174	5.215	5.110	5.047	5.050	**4.874**
	1000	5.298	5.061	5.100	5.073	5.028	5.113	5.019	**4.916**
Goofspiel-4	10	1.295	1.395	1.112	1.280	1.176	1.408	1.151	**1.048**
	100	1.261	1.404	1.136	1.334	1.157	1.400	1.159	**1.024**
	1000	1.258	1.392	1.131	1.401	1.162	1.343	1.135	**1.011**
Goofspiel-5	10	1.853	1.895	1.851	1.889	1.853	1.843	1.815	**1.701**
	100	1.849	1.895	1.862	1.892	1.858	1.842	1.806	**1.704**
	1000	1.843	1.902	1.862	1.881	1.838	1.859	1.804	**1.715**
dark hex (2×2)	10	0.403	0.628	0.540	0.533	0.379	0.449	**0.353**	0.742
	100	0.394	0.628	0.625	0.579	0.348	0.437	**0.336**	0.735
	1000	0.385	0.628	0.568	0.556	0.402	0.408	**0.336**	0.732
liar's dice (1d4)	10	1.406	1.431	1.424	1.442	1.383	1.318	1.352	**1.302**
	100	1.398	1.482	1.452	1.447	1.422	**1.295**	1.369	1.306
	1000	1.410	1.497	1.454	1.443	1.445	**1.288**	1.379	1.308

Temperature scaling. Figure 4b and Table 20 give the full measurement behind Section 7. In the left panel of Figure 4b, the shaded band on the kuhn curve is the per-seed range of $\tau^\star$, which at the highest noise still spans two decades; the leduc markers are open because its $\tau^\star$ is pinned at the grid edge. Online, $\tau = 0.02$ gives exploitability 0.11 against 1.23 at $\tau = 0.5$ at $n_s = 100$; at $n_s = 1$ every τ gives 1.2–1.5, so a small temperature pays only when the budget is large enough to estimate the meta-game.

F Implementation notes

Computing the anchored QRE. We compute $\mathrm{QRE}_{\tau,\rho}$ by magnetic mirror descent (Sokota et al., 2023a) with an *optimistic* (extrapolated) gradient (Rakhlin & Sridharan, 2013), $2g_t - g_{t-1}$: for temperatures much smaller than the step size, plain magnetic mirror descent (MMD) cycles on the nearly bilinear saddle, while the optimistic correction decouples the admissible step size from τ;

Table 16: Paired AUC difference of the anchored QRE against two comparators fixed in advance (mean and paired-bootstrap interval; negative = anchored QRE better): RRD, the strongest replicator baseline, and NE+U, the ε-mixed Nash of Corollary A.9. Bold intervals exclude zero (cellwise, uncorrected); ^{H} marks cells whose paired t-test survives Holm correction at level 0.05 over the 21 cells of the column.

domain	n	vs. RRD	vs. NE+U
matrix (30×30)	10	$\mathbf{-0.187}$ $[\mathbf{-0.322}, \mathbf{-0.037}]$	$\mathbf{-0.331}$ $[\mathbf{-0.495}, \mathbf{-0.149}]^{\mathrm{H}}$
	100	$\mathbf{-0.188}$ $[\mathbf{-0.317}, \mathbf{-0.051}]$	$\mathbf{-0.281}$ $[\mathbf{-0.392}, \mathbf{-0.166}]^{\mathrm{H}}$
	1000	$\mathbf{-0.180}$ $[\mathbf{-0.291}, \mathbf{-0.060}]$	$\mathbf{-0.239}$ $[\mathbf{-0.355}, \mathbf{-0.132}]^{\mathrm{H}}$
Kuhn poker	10	$\mathbf{+0.034}$ $[\mathbf{+0.005}, \mathbf{+0.064}]$	$+0.002$ $[-0.029, +0.035]$
	100	$\mathbf{+0.085}$ $[\mathbf{+0.054}, \mathbf{+0.114}]^{\mathrm{H}}$	$+0.025$ $[-0.015, +0.061]$
	1000	$\mathbf{+0.085}$ $[\mathbf{+0.063}, \mathbf{+0.108}]^{\mathrm{H}}$	$\mathbf{+0.045}$ $[\mathbf{+0.024}, \mathbf{+0.066}]^{\mathrm{H}}$
Leduc poker	10	$\mathbf{-0.297}$ $[\mathbf{-0.417}, \mathbf{-0.164}]^{\mathrm{H}}$	$\mathbf{-0.250}$ $[\mathbf{-0.380}, \mathbf{-0.124}]^{\mathrm{H}}$
	100	$\mathbf{-0.236}$ $[\mathbf{-0.387}, \mathbf{-0.116}]$	$\mathbf{-0.251}$ $[\mathbf{-0.380}, \mathbf{-0.142}]^{\mathrm{H}}$
	1000	$\mathbf{-0.112}$ $[\mathbf{-0.199}, \mathbf{-0.032}]$	$\mathbf{-0.145}$ $[\mathbf{-0.219}, \mathbf{-0.073}]^{\mathrm{H}}$
Goofspiel-4	10	$\mathbf{-0.128}$ $[\mathbf{-0.161}, \mathbf{-0.093}]^{\mathrm{H}}$	$\mathbf{-0.347}$ $[\mathbf{-0.385}, \mathbf{-0.294}]^{\mathrm{H}}$
	100	$\mathbf{-0.133}$ $[\mathbf{-0.163}, \mathbf{-0.104}]^{\mathrm{H}}$	$\mathbf{-0.380}$ $[\mathbf{-0.409}, \mathbf{-0.355}]^{\mathrm{H}}$
	1000	$\mathbf{-0.151}$ $[\mathbf{-0.176}, \mathbf{-0.129}]^{\mathrm{H}}$	$\mathbf{-0.382}$ $[\mathbf{-0.416}, \mathbf{-0.349}]^{\mathrm{H}}$
Goofspiel-5	10	$\mathbf{-0.152}$ $[\mathbf{-0.173}, \mathbf{-0.131}]^{\mathrm{H}}$	$\mathbf{-0.194}$ $[\mathbf{-0.215}, \mathbf{-0.174}]^{\mathrm{H}}$
	100	$\mathbf{-0.154}$ $[\mathbf{-0.176}, \mathbf{-0.133}]^{\mathrm{H}}$	$\mathbf{-0.191}$ $[\mathbf{-0.203}, \mathbf{-0.180}]^{\mathrm{H}}$
	1000	$\mathbf{-0.122}$ $[\mathbf{-0.139}, \mathbf{-0.106}]^{\mathrm{H}}$	$\mathbf{-0.187}$ $[\mathbf{-0.194}, \mathbf{-0.179}]^{\mathrm{H}}$
dark hex (2×2)	10	$\mathbf{+0.363}$ $[\mathbf{+0.261}, \mathbf{+0.449}]^{\mathrm{H}}$	$\mathbf{+0.114}$ $[\mathbf{+0.037}, \mathbf{+0.194}]$
	100	$\mathbf{+0.386}$ $[\mathbf{+0.338}, \mathbf{+0.432}]^{\mathrm{H}}$	$\mathbf{+0.106}$ $[\mathbf{+0.063}, \mathbf{+0.148}]^{\mathrm{H}}$
	1000	$\mathbf{+0.330}$ $[\mathbf{+0.263}, \mathbf{+0.407}]^{\mathrm{H}}$	$\mathbf{+0.104}$ $[\mathbf{+0.035}, \mathbf{+0.178}]$
liar's dice (1d4)	10	$\mathbf{-0.081}$ $[\mathbf{-0.112}, \mathbf{-0.048}]^{\mathrm{H}}$	$\mathbf{-0.128}$ $[\mathbf{-0.161}, \mathbf{-0.093}]^{\mathrm{H}}$
	100	$\mathbf{-0.116}$ $[\mathbf{-0.130}, \mathbf{-0.102}]^{\mathrm{H}}$	$\mathbf{-0.177}$ $[\mathbf{-0.194}, \mathbf{-0.159}]^{\mathrm{H}}$
	1000	$\mathbf{-0.138}$ $[\mathbf{-0.162}, \mathbf{-0.114}]^{\mathrm{H}}$	$\mathbf{-0.189}$ $[\mathbf{-0.215}, \mathbf{-0.159}]^{\mathrm{H}}$

in our experiments the iteration reaches fixed-point residuals below 10^{-8} even at $\tau = 5 \cdot 10^{-3}$. Termination is checked against the logit fixed point (3).

Zero anchor mass. Our implementation floors $\log \rho$ at $\log(10^{-300})$, so the $\varepsilon = 0$ arm instantiates a finite-precision *near*-deadlock: a strategy with zero anchor mass receives solution mass below 10^{-296} (Section D), the finite-precision counterpart of the zero-support mechanism of Theorem 5.1.

Baseline meta-solvers. All solvers act on the same estimated matrix $\hat{A}^t$ and none is tuned per domain. NE: maximin LP. PRD: OpenSpiel's projected replicator dynamics (Lanctot et al., 2019; 2017), 50,000 steps of size 10^{-3} with exploration floor 10^{-8}, output averaged over the last 10% of iterates. RRD: time-averaged replicator dynamics with step 0.05, stopped once the restricted regret is at most $\lambda = 0.05$ (at most 200,000 steps). α-Rank: OpenSpiel's implementation with $\alpha = 10$, the joint stationary distribution marginalized per player. NE+U: the LP Nash mixed with $\varepsilon = 10^{-2}$ uniform (Corollary A.9). MAXENT: the maximum-entropy Nash equilibrium (Balduzzi et al., 2018). QRE-U: the QRE with a fixed uniform anchor and the anchored QRE's schedule. AQRE: Algorithm 1 with $\tau_0 = 0.5$, decay 0.8 per round, $\tau_{\min} = 0.02$, $\varepsilon = 10^{-2}$. The controlled-noise study runs a fixed number of rounds per domain (listed below) with no early stopping; AUC is the trapezoidal integral of exact exploitability (Li & Huang, 2026a; Li et al., 2026a) against $\log_{10}$ cumulative simulations, divided by the log-range. The log-range is each trajectory's own. A solver whose best response is already in the population appends nothing that round, so at a fixed round count the cumulative simulations differ across solvers: at $n = 100$ the seed-averaged final counts span a factor 4.2 on dark hex (400 to 1660) and 3.8 on the matrix (10,820 for α-Rank to 41,110 for NE+U), 1.7 on Goofspiel-4, 1.6 on Kuhn, and coincide on Leduc, Goofspiel-5 and liar's dice. The statistic is

Table 17: Common-budget robustness check. Every (domain, n, seed) cell is re-scored with all eight trajectories truncated at the smallest final cumulative-simulation count among them (log-linear interpolation at the cut) and normalized by that shared log-range. "ratio": mean over seeds of the largest to the smallest final budget among the eight solvers. "best": solver with the lowest mean AUC on the shared axis; where it differs from Tables 15, the trajectory-own best is in parentheses (dark hex only, where the anchored QRE is last on both axes). Last two columns: paired difference of the anchored QRE against RRD and NE+U on the shared axis, 95% paired bootstrap interval, ^{H} = Holm-significant over the 21 cells of the column. The anchored QRE has the lowest mean in 12 of 21 cells on both axes; against RRD it is Holm-significantly better in 12 and worse in 5 cells (trajectory-own: 10 and 5), against NE+U in 15 and 5 (15 and 2), and no Holm-significant cell changes sign.

domain	n	ratio	best	AQRE − RRD	AQRE − NE+U
matrix	10	5.18	AQRE	$-0.301\,[-0.491, -0.089]$	$-0.383\,[-0.570, -0.192]^{\mathrm{H}}$
matrix	100	5.94	AQRE	$-0.289\,[-0.476, -0.115]$	$-0.373\,[-0.538, -0.209]^{\mathrm{H}}$
matrix	1000	4.35	AQRE	$-0.281\,[-0.413, -0.135]^{\mathrm{H}}$	$-0.348\,[-0.477, -0.219]^{\mathrm{H}}$
Kuhn	10	2.75	QRE-U	$+0.042\,[+0.012, +0.075]$	$+0.036\,[+0.006, +0.069]$
Kuhn	100	2.14	QRE-U	$+0.070\,[+0.056, +0.084]^{\mathrm{H}}$	$+0.044\,[+0.028, +0.059]^{\mathrm{H}}$
Kuhn	1000	2.04	QRE-U	$+0.073\,[+0.062, +0.087]^{\mathrm{H}}$	$+0.046\,[+0.036, +0.055]^{\mathrm{H}}$
Leduc	10	1.14	α-Rank	$-0.306\,[-0.428, -0.172]^{\mathrm{H}}$	$-0.267\,[-0.398, -0.139]^{\mathrm{H}}$
Leduc	100	1.03	AQRE	$-0.244\,[-0.393, -0.125]^{\mathrm{H}}$	$-0.257\,[-0.386, -0.147]^{\mathrm{H}}$
Leduc	1000	1.05	AQRE	$-0.112\,[-0.197, -0.033]$	$-0.146\,[-0.219, -0.074]^{\mathrm{H}}$
Goofspiel-4	10	2.03	AQRE	$-0.152\,[-0.195, -0.108]^{\mathrm{H}}$	$-0.328\,[-0.359, -0.287]^{\mathrm{H}}$
Goofspiel-4	100	2.03	AQRE	$-0.162\,[-0.193, -0.132]^{\mathrm{H}}$	$-0.351\,[-0.364, -0.336]^{\mathrm{H}}$
Goofspiel-4	1000	2.00	AQRE	$-0.168\,[-0.191, -0.147]^{\mathrm{H}}$	$-0.349\,[-0.355, -0.343]^{\mathrm{H}}$
Goofspiel-5	10	1.01	AQRE	$-0.152\,[-0.174, -0.132]^{\mathrm{H}}$	$-0.194\,[-0.215, -0.175]^{\mathrm{H}}$
Goofspiel-5	100	1.00	AQRE	$-0.154\,[-0.176, -0.133]^{\mathrm{H}}$	$-0.191\,[-0.203, -0.180]^{\mathrm{H}}$
Goofspiel-5	1000	1.00	AQRE	$-0.122\,[-0.139, -0.106]^{\mathrm{H}}$	$-0.187\,[-0.194, -0.179]^{\mathrm{H}}$
dark hex	10	4.83	NE (QRE-U)	$+0.542\,[+0.511, +0.565]^{\mathrm{H}}$	$+0.560\,[+0.528, +0.586]^{\mathrm{H}}$
dark hex	100	4.53	NE (QRE-U)	$+0.533\,[+0.528, +0.538]^{\mathrm{H}}$	$+0.554\,[+0.548, +0.558]^{\mathrm{H}}$
dark hex	1000	4.22	NE (QRE-U)	$+0.535\,[+0.531, +0.539]^{\mathrm{H}}$	$+0.556\,[+0.552, +0.559]^{\mathrm{H}}$
liar's dice	10	1.01	AQRE	$-0.081\,[-0.113, -0.048]^{\mathrm{H}}$	$-0.129\,[-0.161, -0.093]^{\mathrm{H}}$
liar's dice	100	1.00	α-Rank	$-0.116\,[-0.130, -0.102]^{\mathrm{H}}$	$-0.177\,[-0.194, -0.159]^{\mathrm{H}}$
liar's dice	1000	1.00	α-Rank	$-0.138\,[-0.162, -0.114]^{\mathrm{H}}$	$-0.189\,[-0.215, -0.159]^{\mathrm{H}}$

Table 18: Final exact exploitability (mean over seeds) after the fixed round budget on the two losing domains, all eight solvers. Every solver reaches zero on dark hex except RRD and NE+U; on Kuhn the anchored QRE trails the best solver by at most 0.035, so the AUC deficit of Table 15 is one of speed.

domain	n	NE	NE+U	MAXENT	PRD	RRD	α-Rank	QRE-U	AQRE
Kuhn poker	10	0.428	0.412	0.428	0.410	0.333	**0.331**	0.366	0.352
	100	0.144	0.140	0.138	0.261	**0.115**	0.133	0.122	0.148
	1000	0.051	0.051	0.051	0.131	0.078	0.096	**0.037**	0.055
dark hex (2×2)	10	**0.000**	0.006	**0.000**	**0.000**	0.044	**0.000**	**0.000**	**0.000**
	100	**0.000**	0.006	**0.000**	**0.000**	0.039	**0.000**	**0.000**	**0.000**
	1000	**0.000**	0.006	**0.000**	**0.000**	0.044	**0.000**	**0.000**	**0.000**

therefore a trajectory-averaged exploitability along each solver's own log-budget axis, and the tables compare solvers at equal round counts, not at equal simulation budgets; Table 17 recomputes it on a shared budget axis.

Observation model for matrix domains. A deterministic matrix matchup returns its entry exactly, and repeating it adds no noise, so the matrix rows of Section 7 treat each "episode" as a fresh stochastic observation of the entry under one of two generators. The *additive* model serves the controlled-noise Pareto, the fresh-noise anchor study (Appendix D), and the variance probe of Section 7: the true matrix A has i.i.d. $\mathcal{N}(0,1)$ entries (30×30 for the Pareto and variance-probe matrices, 60×60 in the anchor study), each simulation of entry (x, y) returns $A_{xy} + \xi$ with $\xi \sim \mathcal{N}(0,1)$ i.i.d., and the entry estimate is the mean of n such simulations (unbiased, variance $1/n$). Because this noise is unbounded, no Hoeffding event governs these runs; they support only the relative-solver comparisons of Section 7, never a finite-sample theorem. The controlled-noise Pareto draws each entry's estimate once, the first

Table 19: Adoption of new strategies on the two losing domains: the meta-solution mass placed on the best responses appended in the previous round, pooled over the rounds at which the player did append one (mean, median in parentheses; 10 seeds, same runs as Table 15 re-executed with weight tracing). The final column gives how many seeds favor the anchored QRE's per-run mean over QRE-U's.

domain	n	AQRE	QRE-U	ratio	seeds
Kuhn poker	10	0.141 (0.034)	0.525 (0.518)	3.7	10/10
	100	0.083 (0.034)	0.447 (0.415)	5.4	10/10
	1000	0.078 (0.049)	0.442 (0.396)	5.7	10/10
dark hex (2×2)	10	0.026 (0.003)	0.453 (0.336)	17.4	10/10
	100	0.024 (0.003)	0.435 (0.333)	18.2	10/10
	1000	0.027 (0.004)	0.458 (0.334)	16.7	10/10

Table 20: Scaling exponent $\beta = d\log\tau^\star / d\log\hat{\sigma}$ by matrix. The $\sqrt{\kappa_n}$ heuristic of Remark 6.3 predicts $\beta = 0.5$; five of the six synthetic intervals exclude 0.5 (skill$_2$ contains it). Intervals are 95% paired bootstrap over the ten noise seeds ($B = 200$ resamples, each re-estimating $\tau^\star$ and β and discarded unless ≥ 3 internal levels survive; the retained count is in parentheses); $\pm$ on KUHN-MIX is the half-width, its asymmetric interval symmetrized to the larger side.

matrix	β	interval (retained)	excludes 0.5?
Synthetic full-rank (20 × 20)			
gauss$_0$	0.91	[0.74, 1.06] (198)	yes
gauss$_1$	1.18	[1.04, 1.32] (194)	yes
gauss$_2$	1.04	[0.78, 1.32] (200)	yes
skill$_0$	1.14	[0.70, 1.75] (194)	yes
skill$_1$	0.87	[0.83, 1.18] (200)	yes
skill$_2$	0.74	[0.30, 1.32] (200)	no
Real counterfactual-regret-minimization (CFR) meta-games			
KUHN-MIX (rank 5)	0.00 ± 0.90	flat basin (117)	no
LEDUC-MIX (rank 12)	n.a. (truncated)	pooled only (48)	no

round the entry appears in the population, and reuses it thereafter, so that entry's error is fixed by the n episodes behind it; the fresh-noise studies instead re-sample every currently present entry each round. The *bounded* model serves the terminal-bound experiment (Appendix D) and matches (11): entries are uniform on $[-0.9, 0.9]$, each simulation is a $\{\pm 1\}$ draw with $\Pr(+1) = (A_{xy} + 1)/2$ (expectation A_{xy}), so the estimate is an average of n i.i.d. simulations taking values in $[-B, B]$ with $B = 1$, and every currently present entry is re-sampled fresh each round exactly as Corollary 6.2 states; the κ_n used in its table is $B\sqrt{(2/n)\log(2n_1n_2R/\delta)}$ evaluated at the round-t population sizes and confidence δ/R per round ($R = 2 \cdot 30 = 60$ rounds, $\delta = 0.1$).

Pre-registered experimental protocol. All pass/fail criteria below were fixed in writing before the corresponding runs and evaluated exactly as written; none was modified after seeing results. The written registration will be released together with the code.

Escape study (Section 7). Groups A–E as defined there; $k = 4, \ldots, 14$; deterministic groups (A, B, C) use 3 seeds, stochastic groups (D, E) use 10; round cap 1500, which applies to the simulated groups (B–E); group A replays the adversarial schedule of Zhang & Sandholm (2024) analytically (no sampling), so it is exempt from the cap and its rounds are $2^{k-1} - 1$ by construction; convergence threshold $\max_i \mathrm{gap}_i < 2/k - 10^{-9}$ (the strict margin matters: the adversarial schedule sits exactly at $2/k$). Groups B and D run the same annealed QRE with uniform anchors, $\tau_t = \max(0.25/k, 2/(t+1))$ ($\tau_0 = 2.0$), and differ only in the best-response tie-break (adversarial for B, random for D); at the rounds where these groups escape ($t \approx 3$–5) the anneal exceeds the sufficient cap (7) by roughly an order of magnitude, so the escape is reported as a robustness result beyond the theorem's coverage. Group C's "uniquified" tie-break resolves every state's argmax deterministically by a seed-fixed 10^{-6}-scale perturbation of the payoffs, so its three seeds can differ in which strategy is appended

and hence in the rounds to the gap. Pre-registered criteria: group A must reproduce the exponential ($\log_2$ slope ≥ 0.8 over k, achieved: 1.014, exact match with $2^{k-1}-1$ at every k, all best responses verified legal), and group D must need at most $100k$ rounds at $k = 14$ (achieved: 6). The group-E reading was registered in advance: the theorem's guarantee is for adversarial ties, and the escape claimed here is the *provable* one.

Delaying adversaries (Section 7). Groups F/G/I (poison-Hamming, poison-virgin, random; 5 seeds) and H (adversarial; 3 seeds) under the anchored-QRE meta-solver with the temperature capped per (7); $k = 4, \ldots, 14$, cap 1500. Pre-registered criteria: (i) every run must respect the $4k$ bound of Theorem 4.3 (a violation would indicate a bug in proof or code; none occurred, maximum 4 rounds), and (ii) the growth of max rounds in k is reported against constant, $c\sqrt{k}$, and ck fits, with the remark wording bound to the best fit (constant).

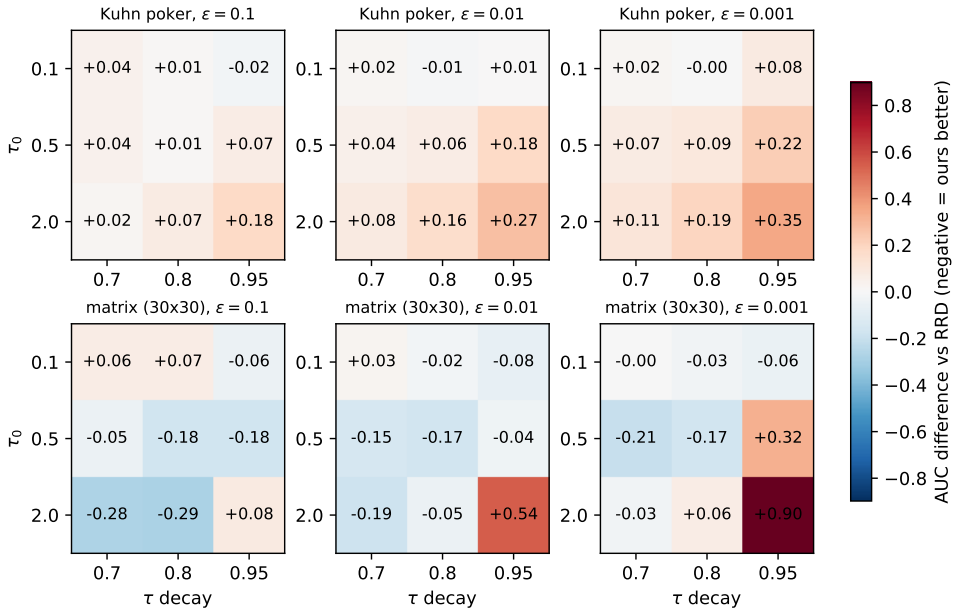
Controlled-noise Pareto (Section 7). Seven domains (Kuhn poker (Kuhn, 1950), Leduc poker (Southey et al., 2005), Goofspiel (Ross, 1971), dark hex and liar's dice in their OpenSpiel implementations (Lanctot et al., 2019), and a random matrix); 10 seeds each. The three heavier domains (Leduc, liar's dice, Goofspiel-5) were pre-registered at 5 seeds; seeds 5–9 were appended after review, and every table and verdict below is recomputed on the pooled 10 (at 5 seeds the never-worse criterion held on $5/7$ domains, the two cells that turned significantly worse being Leduc at $n = 10$ and liar's dice at $n = 1000$). PSRO rounds per domain fixed in advance (25 matrix, 15 Kuhn, 12 Leduc and liar's dice, 16 Goofspiel-4 and dark hex, 10 Goofspiel-5); $n \in \{10, 100, 1000\}$ episodes per meta-game entry; metric: exact exploitability AUC over $\log_{10}$ cumulative simulations, paired per seed against the strongest baseline with a $2{,}000$-resample bootstrap; intervals are cellwise and uncorrected for the per-cell baseline choice, so they are descriptive, not simultaneous; the pre-registered "significantly worse/win" wording below refers to these cellwise tests. Pre-registered criteria: a majority of domains never significantly worse, and at least three domains with a significant win at some noise level (achieved: $3/7$, failed, and 4, met). The fixed-comparator table adds a Holm correction over its 21 cells per column, a post hoc addition.

Hyperparameter ablation (Section 7). Full grid $\tau_0 \in \{0.1, 0.5, 2.0\} \times$ decay $\in \{0.7, 0.8, 0.95\} \times \varepsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ ($\tau_{\min} = 0.02$ fixed), $n = 100$, 5 seeds, on Kuhn and the random matrix, paired against the RRD runs of the main study (same seeds, not re-run). Pre-registered criteria: (i) existence of a Kuhn combination matching RRD, (ii) monotonicity of AUC in ε, both reported as-is; (iii) robustness required the matrix win to persist on a majority of combinations; this criterion *failed* ($9/27$).

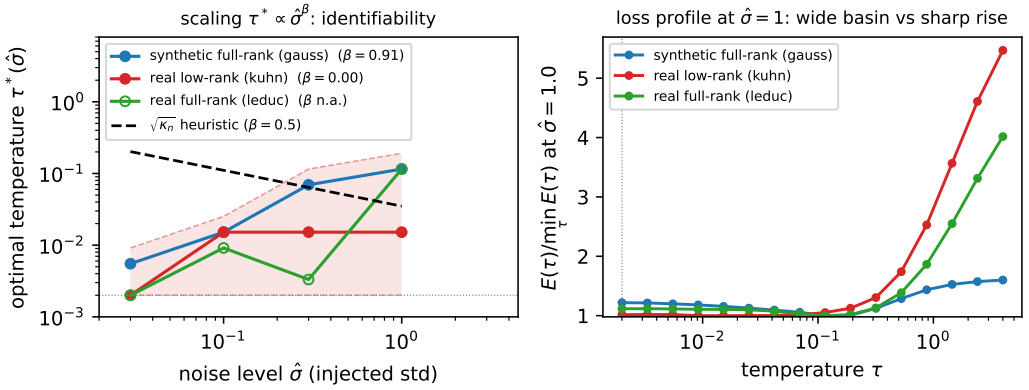
Deadlock and repair (Section 7). 60×60 Gaussian matrix PSRO, anchor = previous meta-solution, 10 seeds, $\varepsilon \in \{0, 10^{-1}, 10^{-2}, 10^{-3}\}$. Hard deadlock := stalled exploitability and maximum new-strategy weight below 10^{-6} after round 5. Pre-registered criteria: the raw-anchor arm deadlocks on all seeds and every ε arm on none (achieved: $10/10$ and $0/10$ each).

Temperature scaling (Section 7). Eight meta-game matrices: six synthetic full-rank (20×20 Gaussian and skill games); KUHN-MIX, the 12×12 meta-game of a Kuhn CFR strategy pool of rank 5, whose three strategy-pool seeds produce identical pools because CFR is deterministic (the ten noise seeds are independent); and LEDUC-MIX, a 12×12 Leduc pool built with the C++ CFR solver (Li & Huang, 2026b; Li et al., 2026b) at 20000 iterations. Noise $\hat{\sigma} \in \{0.03, 0.1, 0.3, 1.0\}$; τ grid log-spaced on $[0.002, 4.0]$ with 16 points; 10 noise seeds; the anchored QRE is solved by optimistic mirror descent (50000 iterations, uniform anchors). With $E(\tau, \hat{\sigma}) = \frac{1}{10}\sum_s \mathrm{NashConv}_A(\mathrm{QRE}(\tau, A + \hat{\sigma}N_s))$, we set $\tau^\star(\hat{\sigma}) = \arg\min_\tau E(\tau, \hat{\sigma})$; a truncation audit excludes $\tau^\star$ outside $[\mathrm{TAUS}[1], \mathrm{TAUS}[-2]]$, and a matrix is retained only if ≥ 3 internal noise levels remain. The point β is a log–log regression over internal levels, and the 95% interval comes from $B = 200$ bootstrap resamples over noise seeds, each of which re-estimates $\tau^\star$ and β and is discarded unless ≥ 3 internal levels survive.

Pre-registered criteria and outcomes: (G0) truncation and monotonicity reported per matrix; done, with KUHN-MIX truncated at $\hat{\sigma} = 0.03$ and LEDUC-MIX at $\hat{\sigma} \in \{0.03, 0.3\}$. (G1) The real class against 0.5: the pooled bootstrap interval across the real class (KUHN-MIX and LEDUC-MIX, $n = 165$ retained bootstrap samples, 117 Kuhn and 48 Leduc) is 0.52 ± 0.81, which contains 0.5, but the valley is flat, within 2% of the minimum across a $34\times$ range of τ, with per-seed $\tau^\star$ spanning two decades at the highest noise, so the $\sqrt{\kappa_n}$ heuristic is *neither validated nor refuted* and the trade-off degeneracy is the main conclusion.



(a) Hyperparameter ablation.



(b) Temperature scaling.

Figure 4: (a) Paired AUC difference against RRD at $n = 100$ (negative is better): on Kuhn the schedule staying hot degrades performance, the bias term in action; on the matrix wins persist outside the bias-dominated corner. (b) Left panel: $\tau^\star$ against noise $\hat{\sigma}$ (log–log); the plotted synthetic matrix (gauss_0) moves with $\beta = 0.91$ (class mean 0.98, Table 20), far steeper than the $\sqrt{\kappa_n}$ heuristic's 0.5, while real games are flat or pinned at the grid edge. Right panel: the loss $E(\tau)$ relative to its minimum at $\hat{\sigma} = 1$; the synthetic matrix has a wide basin, the real games a sharp rise once τ passes ≈ 0.3.

(G2) Synthetic class deviation from 0.5: β mean 0.98, range 0.74–1.18, with $5/6$ intervals excluding 0.5, steeper than square-root scaling (each interval is listed in Table 20). (G3) Mechanism: the bias local exponent a on $\tau \in [0.02, 0.1]$ and the variance local exponent b on the $\tau^\star(1.0)$ window. Real games give $a \approx 0.8$ (sub-linear, kuhn 0.79, leduc 0.80) against synthetic $a \approx 1.0$–1.2, and kuhn's variance term barely decays ($b \approx -0.08$), which together produce the flat basin. (G4) Online PSRO loop on a 30×30 Gaussian matrix, fixed $\tau \in \{0.02, 0.05, 0.1, 0.2, 0.5\}$, $n_s \in \{1, 11, 100\}$ episodes per entry, 25 rounds, 3 seeds: $\tau^\star$ decreases with budget ($0.5 \to 0.05 \to 0.02$), confirming the direction, but fewer than three internal noise levels preclude fitting a β. Sensitivity check on a 25-point grid for KUHN-MIX and gauss_0: β changes by at most 0.04, ruling out a discretization artefact.

G Additional related work

Meta-strategy solvers and empirical game-theoretic analysis (EGTA). Alternatives to exact Nash for empirical game-theoretic analysis (Wellman, 2006) include α-Rank (Omidshafiei et al., 2019) and its PSRO instantiation (Muller et al., 2020), structure-exploiting solvers (Konicki et al., 2022), gradient-based Nash approximation (Gemp et al., 2022), and regularized replicator dynamics (Wang & Wellman, 2023), our strongest baseline. Co-learning with estimated world models extends the setting beyond fixed payoffs (Smith & Wellman, 2023). These solvers treat the meta-game as fixed; their behavior under estimation noise is analyzed piecemeal; we use the Lipschitz sensitivity of strongly monotone variational inequalities and carry it through the PSRO loop to a finite-sample bound on terminal exploitability under noisy restricted payoffs.

Concurrent and adjacent work. Zhang et al. (2026) prove, for two-player zero-sum normal-form games, a lower bound linear in the number of pure strategies for any meta-solver that reads only the restricted payoff matrix, and propose an oracle that looks outside the restricted game. Our result is orthogonal: it fixes the extensive-form family of Zhang & Sandholm (2024), keeps the meta-solver restricted-game-based, and shows that perturbing it already removes the exponential worst case there; the two statements are compatible since their hard instances differ. The POSG construction of Zhang & Sandholm (2024), whose selections are unique throughout, is likewise outside the reach of selection perturbation (Remark A.15). The term *KL-anchored* appears in Farina et al. (2019), where the anchor is a fixed reference strategy; our anchor is adaptive, the previous meta-solution mixed with an ε-uniform distribution, and it is the adaptivity that drives the deadlock analysis of Theorem 5.1. Mixing the current equilibrium with past equilibria as a *training target* for the best-response learner (Wright et al., 2019) improves stability on the oracle side, whereas we perturb the meta-solver and leave the oracle exact. Finally, the observation that logit-equilibrium meta-strategies behave better than exact Nash on noisy empirical games goes back to Kiekintveld & Wellman (2008); Theorem 6.1 gives that observation an explicit modulus, $2/\tau$, and Corollary 6.2 carries it through the population loop.

H Conclusion

Double-oracle's exponential lower bound and the anchoring deadlock are the same pathology: a meta-solution forced to place zero mass where it is needed, certifying the same $2/k$ gap. An anchored QRE meta-solver under the temperature condition (7), with an ε-uniform mixture, cures both. It collapses the worst case from 2^{k-1} to at most $4k$ rounds under exact QRE solves and best-response oracles, and an ε-mixed Nash inherits the escape with inverse-polynomial oracle tolerance. It restores a mass floor and termination, and it is $(2/\tau)$-Lipschitz where exact Nash is discontinuous. On the adversarial-selection side of the bound, an interior solution removes the selection adversary's lever entirely on this hard family.